Efficient estimation of crash count data via metaheuristic solution algorithm

Zeke Ahern[1], Alexander Paz[1], Paul Corry[2]

[1]School of Civil & Environmental Engineering, Queensland University of Technology, Brisbane, Australia

[2]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia

Email for correspondence: zeke.ahern@hdr.qut.edu.au

Abstract

Efficient estimation of modeling crash-counts is complex, requiring advanced knowledge, experience, ad hoc and time-consuming processes. Challenges involved in developing the model include; identification of significant road factors for model specification, an adequate functional form and the type of model such as Poisson. To overcome these barriers, a symbolic regression technique assisted by the Harmony-Search solution algorithm was developed to identify likely crash contributing factors for Head-on Fatal and Serious Injury (FSI) collisions on the Queensland state road network. The applicability of the approach is demonstrated by comparing it with the results found in Behara et al. (2021). As a result of the solution algorithm, human bias' are minimised since decisions are updated based on the goodness of fit rather than context-specific information. Additionally, time is reduced as more models can be tested more efficiently, and new solutions that are worse than the incumbent solution are considered which enables the escape of local-optima regions that an analyst may not be able to identify.

1. Introduction

Statistical models such as Poisson regression (Arndt, 2004), Negative Binomial or its variants (Yu et al., 2019), are the most common approach for estimating crash counts (Mannering and Bhat, 2014). This is because they are easy to estimate and can be derived to handle over-dispersion in the model. These models only handle linear relationships between features and do not allow parameter estimates to vary across observations; recent studies have emphasised the need to account for unobserved heterogeneity in the data (Malyshkina and Mannering, 2010; Zeng et al., 2017). In these approaches, the distribution of random effects and the selection of potential explanatory variables need to be assumed before estimation. This model specification process relies greatly on human judgement to include context-specific information in the model. Furthermore, it is time consuming, subject to ad hoc trial-and-error approaches, of which local-optima solutions are likely to be proposed. A method that lowers the time and effort required to find an efficient solution while avoiding human bias' to the available information is therefore desirable.

Harmony search (HS) is a metaheuristic optimisation algorithm that is inspired by improvisations made by jazz musicians in seek perfect harmony. Musicians will tweak their pitch in correspondence with each other to obtain the desired harmony, which is analogous to the optimisation process. HS has proved to be effective for feature selection problems (Paz et al., 2019). Advantages of HS include; ability to escape local optimal solutions (Diao and Shen, 2012), quick convergence (Inbarani et al., 2015), and small number of hyperparameters which are easy to tune (Alia and Mandava, 2011; Kattan et al., 2010).

This study proposes a harmony search solution algorithm to find the best model specification by optimising the Bayesian information criterion (BIC). BIC is a criterion for model selection that prevents overfitting by penalising the number of parameters (Markon and Krueger, 2004). The specification includes the most significant explanatory variables and the functional form of the explanatory variables within a Poisson regression model.

2. Methodology

The following notation is used to formulate the problem:

Table 1: Notation and definitions.

Notation	Definitions
I	number of road observations
X	vector of potential explanatory variables, including interactions between road factors
K	number of explanatory variables
S	number of parameters include in the model
α_k	Indicator variable taking value 1 if potential explanatory variables x_k is included; 0 otherwise
T	set of transformations
τ_k	variable to select the transformation from T applied to α_k
β	vector of coefficients for potential explanatory variables

The problem is formalised as a bi-level optimisation problem. The lower level objective function is maximum likelihood estimation (MLE) to find β, constrained by the following decisions:

- α_k: the explanatory variables that specify the model
- τ_k: the transformations on α_k to create the functional form

The upper-level objective function is to minimise Bayes Information Criteria (BIC); expressed as:

$$\text{BIC} = ln(N)S - 2ln(LL)$$

Harmony search is used to solve the symbolic regression problem. The algorithm can be observed in Algorithm 1. The hyper-parameters for HS include: HMS (harmony memory size), $HMCR$ (harmony consideration rate), PAR (pitch adjustment rate). A harmony memory set HM is first generated to the size of HMS with random specifications $\alpha_k, \tau_k \in K$. The generated harmony set is ranked, and improvisations are made based on the set to generate a new solution. The harmony set is updated when a new solution is generated that outperforms an existing solution.

Algorithm 1 Harmony Search

```
Inputs:
    HMS = 30; HMCR = 0.75;PAR = 0.5; iter = 300
    HM = generate random specification
    sort(HM)
for i = 1, 2, .., iter do
    procedure Improvise Harmony(H)
        for j = 1, 2, .., size(H) do
            if rand(0, 1) ≤ HMCR then
                r = rand(1, HMS)                          ▷ Select random specification
                H[j] = HM[r, j]
            else
                H[j] = rand(d(j))                         ▷ Generate random specification
            end if
            if rand(0, 1) ≥ PAR then
                Conduct pitch adjustment
            end if
        end for
        BIC(H)                                            ▷ Evaluate Objective BIC
        while All p_val of features j in H_j ≤ 0.05 do
            Remove insignificant features
            BIC(H)
        end while
        if BIC(H) ≤ BIC(HM[HMS]) then                     ▷ Compare solution to worst in HM
            H = HM[HMS]                                   ▷ Replace worst
            sort(HM)
        end if
    end procedure
end for
```

3. Experiments and Results

A dataset with 1848 observations which detail road factors with crash counts for the Queensland state road network was used in this study. This dataset was initially used by Behara et al. (2021). The number of road factors considered in the model is 20. All road factors paired were considered as potential interactions and included in the set X. The set of possible transformations applied within HS was inputted as $T = [exp(), pow(2), pow(3), ln(), sqrt()]$. The hyperparameters of the HS algorithm are outlined in Algorithm 1.

The results of the proposed approach are comparable to the models developed by Behara et al. (2021) in terms of MLE (see Table 2). The HS solution algorithm was shown to converge with minimal CPU time (141 seconds), as depicted by Figure 1. This is a significant result, as an efficient model was produced quickly, and is comparable to the time-consuming efforts of an analyst. The coefficients of final model can be observed in Table 3.

Table 2: Harmony search results compared

Statistical Models	**HS**	RPNB	RPNB-HET	RPP	RPP-HET
MLE	-286.08	-283.12	-283.3	-274.5	-276

Table 3: Analysis of exposures variables

Response	Mean	Std. Error
Intercept	-10.5407	0.973
pow(RSHS, 2)	1.5240	0.336
pow(AADT, 3)	-18.5658	4.157
pow(MCV, 2)	-52.1395	12.476
AADT:MCV	75.0005	15.470
pow(MCV, 2)	-52.1395	12.476
AADT:MCV	75.0005	15.470

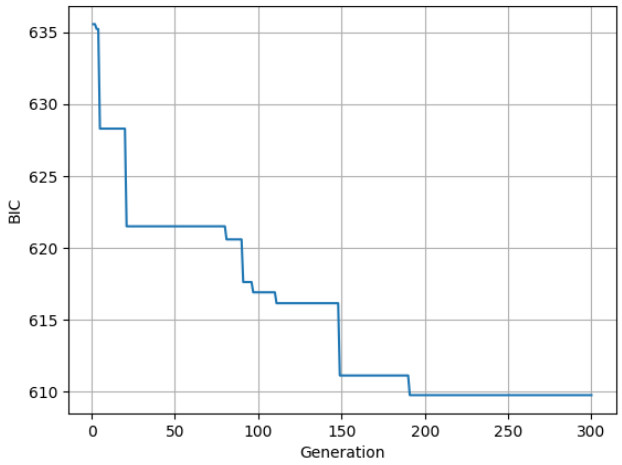

Figure 1. HS convergence

4. Conclusion

The proposed approach achieved the intended aims of eliminating human bias' which have driven model specifications as well as producing efficient solutions in a timely manner. The final model found by the HS algorithm was comparable to the advanced model specifications constructed by Behara et al. (2021). The final model utilised only 141 seconds of CPU time. Further development of more complex functions could be extended within the solution algorithm because CPU time is flexible to facilitate better estimates.

Future work will test more complex model forms (Negative Binomial with Random Parameters and Heterogeneity in the Means), and more transformations on the explanatory variables. These extensions could allow the HS to return more efficient solutions than the results found by the analyst. We also want to explore a clusterwise regression approach, where clusters of sites within each predefined site subtype are created based on the observed crash trends. This may lead to an optimal number of estimation functions, further classifying site subtypes into various subgroups to provide better crash estimates that minimise the overall estimation error (Khadka et al., 2020).

References

Alia, O. M. and Mandava, R. (2011). The variants of the harmony search algorithm: an overview. *Artificial Intelligence Review*, 36(1):49–68.

Arndt, O. K. (2004). *Relationship between unsignalised intersection geometry and accident rates*. PhD thesis.

Behara, K. N., Paz, A., Arndt, O., and Baker, D. (2021). A random parameters with heterogeneity in means and Lindley approach to analyze crash data with excessive zeros: A case study of head-on heavy vehicle crashes in Queensland. *Accident Analysis & Prevention*, 160:106308.

Diao, R. and Shen, Q. (2012). Feature Selection With Harmony Search. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(6):1509–1523.

Inbarani, H. H., Bagyamathi, M., and Azar, A. T. (2015). A novel hybrid feature selection method based on rough set and improved harmony search. *Neural Computing and Applications*, 26(8):1859–1880.

Kattan, A., Abdullah, R., and Salam, R. A. (2010). Harmony Search Based Supervised Training of Artificial Neural Networks.

Khadka, M., Paz, A., and Singh, A. (2020). Generalised clusterwise regression for simultaneous estimation of optimal pavement clusters and performance models. *International Journal of Pavement Engineering*, 21(9):1122–1134.

Malyshkina, N. V. and Mannering, F. L. (2010). Zero-state Markov switching count-data models: An empirical assessment. *Accident Analysis and Prevention*, 42(1):122–130.

Mannering, F. L. and Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic methods in accident research*, 1:1–22.

Markon, K. E. and Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics*, 34(6):593–610.

Paz, A., Arteaga, C., and Cobos, C. (2019). Specification of mixed logit models assisted by an optimization framework. *Sustainability (Switzerland)*, 30:50–60.

Yu, R., Wang, Y., Quddus, M., and Li, J. (2019). A marginalized random effects hurdle negative binomial model for analyzing refined-scale crash frequency data. *Analytic methods in accident research*, 22:100092.

Zeng, Z., Zhu, W., Ke, R., Ash, J., Wang, Y., Xu, J., and Xu, X. (2017). A generalized nonlinear model-based mixed multinomial logit approach for crash data analysis. *Accident Analysis & Prevention*, 99:51–65.