Submission ID: 70

Lund, Sweden
10-12 November 2021
iCSC International Cycling Safety Conference

Changes in interpretation and empirical findings when fully specifying latent classes in a Latent Class Choice Model

Marcus Skyum Myhrmann and Georges Sfeir

Keywords: Injury severity, Single-bicycle crashes, Gaussian-Bernoulli Latent Class Choice Model, Unobserved heterogeneity, Fully specified latent classes

Bicycling provides a sustainable alternative to the car as a mode of transport. It is being heavily endorsed given its environmental and physical health benefits. However, it is also a method of transport that carries significant risk of injury, something that presents a barrier to increase the amount of cycling. Therefore, understanding the circumstances of associated injuries are of utmost importance, if preventive efforts are to be made to increase the cycling mode share.
The methodological frontier in modelling injury severity outcomes has in recent years increasingly focused on addressing unobserved heterogeneity. Latent class choice models (LCCMs) have proven especially adept at this task. They are a type of finite mixture model that address the unobserved heterogeneity arising from unobserved groups assumed to constitute the crash population. Temporary latent classes are constructed during the model estimation process, with the aim of optimizing the ability of predicting the unconditional response. However, the interpretability of the latent classes needs careful and detailed investigation.
This study develops and compares two latent class choice models, the socio-demographic LCCM and the fully specified LCCM. The former uses only socio-demographic variables (age and gender) to describe the assignment to latent classes and the remaining explanatory variables (e.g. weather conditions, road conditions etc.) to estimate the injury severity outcome. The later uses all available explanatory variables in addition to socio-demographic variables to assign individuals to different classes while the injury severity is modeled using only alternative-specific constants. The two approaches make use of a recently developed LCCM framework, Gaussian-Bernoulli Latent Class Choice Model (GBLCCM). In such framework, the probability of an individual belonging to a specific class is modeled using a Gaussian-Bernoulli Mixture Model, which has shown the ability to capture more behavioral heterogeneity than a random utility formulation, while the injury severity outcome is formulated using a multinomial logit regression model.
The models are applied on a single-bicycle crash data obtained through medical records merged with road data collected between 2010 and 2015. The injury severity outcome levels were: 'severe', 'slight', and 'no evident' injury. The representation of the 3 outcomes are approximately 19%, 60%, and 21% for the three levels, respectively.
Results showed that the best trade-off between goodness-of-fit and parsimony is with two and six classes for the socio-demographic LCCM and fully specified LCCM, respectively. The socio-demographic model has better log-likelihood. However the classes are only determined to the extent that "class 1" is associated to younger cyclists and "class 2" to older. The gender separation in both classes is approximately 50/50. Individuals from "class 2" have higher likelihood of severe injury. According to the fully specified model, the second class is associated to an increased baseline likelihood of severe injuries as opposed to the third and fourth classes which have lower likelihood of severe injuries. Around 80% of cyclists from the second class crashed on roads with an average annual daily cyclist traffic (AADCT) of 3000 while this number is only 30% and 0% for the fourth and fifth classes, respectively. Moreover, the third and fourth classes are more likely to include younger cyclists. The remaining classes (first, fifth, and sixth) don't associate to a significant increase or decrease in the baseline likelihood of severe injury. Finally, both models identify elasticities with similar signs but different order of magnitude.
To sum up, the socio-demographic LCCM has better goodness-of-fit measures. However, the latent classes of the fully specified LCCM are more readily interpreted and could as such enable better preventive efforts.