Monsters and mixtures

Richard Torkar

April 2021

OVER-DISPERSED OUTCOMES we've already covered quickly in the previous lecture notes, i.e., Beta-Binomial and Gamma-Poisson (or Negative-Binomial as it is also called). Since they combine several distributions we call them mixture models. We estimate a main parameter (e.g., what we find in Binomial or Poisson) and an auxiliary parameter for the variance (or as we will see next, for the zero-inflation).

ZERO-INFLATED OUTCOMES can be used to model data that has more zeros in the outcome than expected (we model them separately). In the case of over-dispersed outcomes we modeled the dispersion (variance) separately. Several distributions exist with a zero-inflated (zi) part. These are also seen as mixture models.

Often one can see Binomial, Poisson, and Negative-Binomial with a zero-inflated part.

It is worth to note that in some cases the two parameters being estimated can use disparate link functions. Take care and always check!

ORDNUNG MUSS SEIN! Ordered categorical outcomes or predictors are common when conducting surveys. Every year a number of students do a master thesis where there's a survey component that needs to be analyzed. Often the survey consists of Likert scales.

i.e., from a scale 1–5, where 5 is best, how do you rate Richard's course?

There are several ways to model ordered categorical (ordinal) data, but not until quite recently was it possible to use them easily in Bayesian data analysis. Software engineering, generally speaking, handles ordered categorical data by assuming that the conclusions do not depend on if a regression or ordinal model is used. The problem is, of course, that relying on an incorrect outcome distribution will lead to subpar predictive capabilities of the model (Bürkner and Vuorre, 2019). This, in combination with the fact that effect size estimates will be biased when averaging multiple ordinal items, and that data can be non-normal, is something a researcher should want to handle (Liddell and Kruschke, 2018).

Today, we have at least three principled ways to model ordinal data: Adjacent category (Bürkner and Vuorre, 2019), Sequential (Tutz, 1990), and Cumulative models (Walker and Duncan, 1967). These models have been developed and refined in a Bayesian framework mostly because of needs from other disciplines, such as psychology (Bürkner and Vuorre, 2019).

The absolutely most common likelihood to use is the Cumulative, which is also what's covered in the book.

Modeling ordered variables (as outcomes or as predictors) is relatively straightforward. Interpreting the estimated outcome is also relatively easy, but interpreting the predictors' estimates is not as straightforward. I'd like you to spend some time understanding what the estimated values mean when using ordered variables as predictors. Also, ordered variables need a special prior, i.e., Dirichlet, but that one is easy to figure out.[1]

[1] We model each category separately and they should all sum to one.

References

P.-C. Bürkner and M. Vuorre. Ordinal regression models in psychology: A tutorial. *Advances in methods and practices in psychological science*, 2(1):77–101, 2019. DOI: 10.1177/2515245918823199.

T. M. Liddell and J. K. Kruschke. Analyzing ordinal data with metric models: What could possibly go wrong?". *Journal of Experimental Social Psychology"*, 79:328–348, 2018. ISSN 0022-1031. DOI: 10.1016/j.jesp.2018.08.009.

G. Tutz. Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43(1):39–55, 1990. DOI: 10.1111/j.2044-8317.1990.tb00925.x.

S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54 (1–2):167–179, 06 1967. ISSN 0006-3444. DOI: 10.1093/biomet/54.1-2.167.