Multinomial and Dirichlet-multinomial modeling of categorical time series

Erik Thorsén[*]

June 2014

Abstract

This Bachelors thesis considers two categorical time-series regression models, the multinomial logit regression model and the Dirichlet-multinomial regression model. The interesting differences between these two regression models lies in their respective variance-covariance structure where the Dirichlet-multinomial is more flexible. Aim was to present a mathematical description of the two models and compare them for a specific dataset. The two models where fitted, under the assumption of no autocorrelation, for a case analysis on proportions of age-categories of reported rotavirus cases from Brandenburg, Germany. The proportions of age-categories where of interest because one would like to see if there was a age-shift from young children and infants (00-04) to elderly (70+), after the introduction of a vaccination programme for infants and young children in 2009 (Koch and Wiese-Posselt, 2011). Data show strong seasonal variation. A graphical presentation of the two categorical regression models show that they fit mean values almost equal. However, large differences are shown in the goodness-of-fit measures AIC and BIC and in predictive intervals, constructed by sampling. On average, a predictive interval for elderly (70+) using the Dirichlet-multinomial time-series regression contained approximately 90% of the observations while as the multinomial logit regression contained approximately 50-60% of the observations. These results display that when data shows tendencies of over-dispersion, the Dirichlet-multinomial regression model is more adequate to model proportions with than the multinomial logit regression.

[*]Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:ethorsn@gmail.com . Supervisor: Michael Höhle.