University of Toronto
Dalla Lana School of Public Health
CHL 5210 – Categorical Data Analysis
Assignment #2 – due 2016 December 12 at 4:30pm

A hard copy of the completed assignment must be submitted to the Graduate Office, 6th Floor of the Public Health Sciences building by 4:30pm on the 12th of December.

This assignment pertains to a fictional study of individuals who are active patients of a clinic specializing in Post-Traumatic Stress Disorder (PTSD). The study sought to examine patient characteristics associated with self-medication through alcohol consumption. Three characteristics were identified: previous history of alcohol abuse, concurrent severity of PTSD symptoms, and concurrent emotional state. Agreement to participant and written informed consent were obtained from 48 of the patients. During face-to-face interviews with the patients, questionnaires were completed to quantify consumption of alcohol, severity of PTSD symptoms, and emotional state over the previous 24 hours. History of alcohol abuse was extracted from their health record at the clinic.

The data for this assignment are contained in the file "assignment2_data". There are 48 observations on the following five variables:

- *id*: the participant's unique study identification number;
- *etoh*: the number of drinks consumed by the participant over the past 24 hours;
- *hx*: indicator of whether the participant has a history of alcohol abuse (hx = 1) or not (hx = 0);
- *ptsdsx*: score indicating severity of PTSD symptoms experienced by the participant over the past 24 hours (greater severity indicated by higher values);
- *affect*: score indicating the participant's emotional state over the previous 24 hours (enthusiasm and alertness giving rise to positive values and lethargy and sadness giving rise to negative values).

Consider that you have been asked by one of the investigators to provide your impression of the data. She will be incorporating your feedback into a written manuscript and a conference talk which you will not be attending. Assume that the investigator is very intelligent but is not familiar with statistical jargon. That is, you will need to explain very clearly what the investigator needs to know in order to defend their conclusions.

Structure your report to address the following items.

1. Describe the rationale for accommodating "excess zeros" in the model. Cover both how these data indicate the necessity for such accommodation and why this indication should not be ignored. [8 points]
2. Fit and interpret a hurdle model (with Poisson count distribution) for *etoh* that contains main effects of *hx*, *ptsdsx*, and *affect*. Do not include interaction terms in the model. Note, the pscl package in R contains a hurdle model fitting function. [12 points]
3. The investigator would like to know if any of these three explanatory variables are superfluous. Please address this query. [4 points]
4. The investigator suggested that they have had an affinity for using a "forward selection" model building approach that identifies, at each step, candidate main-effect variables for which the Wald test has a p-value < 0.25. Apply this model-building algorithm and describe the fitted final model. How does interpretation of the associations differ between this fitted final model and the fitted model from Question 2 (above). Which model would you recommend as a more appropriate representation of the data? [8 points]
5. The investigator is curious about the use of a hurdle model to accommodate excess zeros. Explain the difference between hurdle and zero-inflated models. Explain what you need to know about this study in order for you and the investigator to decide whether the hurdle or zero-inflated model is more appropriate for this study. [8 points]