PHST 684
Categorical Data Analysis
Spring 2023
Due: 04/30/2023 11:59 pm EST

Final Project

The data is from a study, in which, an advanced two-level Road Departure Avoidance System at Horizontal Curves (RDAS-HC) has been developed with Connected Vehicle technologies and In-Vehicle Heads-Up Display in a simulation environment. The participants finished a horizontal curves driving test under various circumstances such as low visibility and wet pavement conditions. Researchers would like to examine whether the RDAS-HC system can reduce the probability of lane departures in horizontal curves and whether the RDAS-HC system distracts drivers.

The data sets are posted to Blackboard as data1.dat, data2.dat, and data3.dat. The following variables are contained in those data sets.

- **Distraction**: the response of participant to the question whether RDAS-HC system distracts the driver. It has values from -3 to 3. -3 denotes no distraction at all, 0 notes neutral, and 3 denotes significant distraction.
- **LD**: lane occurrence (0: no; 1: yes)
- **Level**: the level of RDAS-HC system (0: control; 1: level 1 of RDAS-HC; 2: level 2 of RDAS-HC)
- **Male**: the gender of participant (0: female; 1: male)
- **Night**: the illumination condition (0: daylight; 1: night)
- **Wet**: the pavement condition (0:dry; 1: wet)

Analyze these data using the methods discussed below. Provide a typed write-up of your results, and be sure to include tables and figures to support your analysis. Include your code and any relevant output (edited, not gratuitous!) in your write-up. Do include summary tables (in a nice format) in the body of your text (e.g. tables with parameter values and standard errors, deviance values, etc.). Your analysis should address each of the items listed below.

1. Use the following codes to obtain a table.

```
Data1 <- read.table("Data1.dat",header = TRUE)
Tab1 <- aggregate(Data1$Count,
        by=list(Level=Data1$Level, Male=Data1$Male, LD=Data1$LD),
        FUN=sum)
names(Tab1)[4]="Count"
```

(a) Reorganize the data to get a $3 \times 2 \times 2$ contingency table containing **Level**, **Male**, and **LD** and present the table using functions xtabs and ftable. [3 pts]

(b) Calculate the observed conditional odds ratios giving the effectiveness of reducing lane departure occurrence for males compared to females given each level of RDAS-HC system and compare the conditional odds ratios to the marginal odds ratio between **Male** and **LD**. Present these using a nice table (Hint: you may need to use the tapply function). [7 pts]

(c) Use the Cochran-Maentel-Haenszel method to test for conditional independence of gender and lane departure, given the level of RDAS-HC. [5 pts]

(d) Give a 95% confidence interval for the common odds ratio based on the CMH method. [5 pts]

2. Use data1.dat and treat **LD** as the response variable and other variables as predictors. (**Pease just use data1.dat and do not run the codes in Problem 1**).

(a) Reorganize the data so that the resulting data frame contains two columns that contain the counts of **LD** yes and **LD** no, respectively. [3 pts]

(b) Fit a saturated logit model and use step function to select a model. [7 pts]

(c) Obtain the standardized Pearson residuals from the model selected in part (b). What do you conclude? [5 pts]

(d) Test whether the association between RDAS-HC level and lane departure is homogeneous, given the other variables. (Hint: which logit model implies the homogeneous association). [5 pts]

(e) Use likelihood ratio test to test for conditional independence of RDAS-HC and response, given the other variables. Be sure to clearly state the model and hypothesis you are testing. [5 pts]

Bonus: Use the Wald test to test for conditional independence of RDAS-HC and response, given the other variables. [5 pts]

(f) Draw the ROC curve for the model selected in Part (b). You may use or modify the following codes to create a suitable data structure. Let Tab2 be the data frame you created in (a) and the column names of LD Yes and No be Yes and No. I assume that they are columns 5 and 6. If not, you can modify the codes.

```
Tab3.1=(Tab2[rep(seq_len(nrow(Tab2)), Tab2$No),])[,-c(5,6)]

Tab3.1$y = 0

Tab3.2=(Tab2[rep(seq_len(nrow(Tab2)), Tab2$Yes),])[,-c(5,6)]

Tab3.2$y = 1

Tab3=rbind(Tab3.1, Tab3.2)
```

Fit Tab3 with the model you selected in Part 2(b) and produce the ROC curve.

3. Use data2.dat and treat **Distraction** as a nominal response variable and other variables as predictors.

   (a) Fit a multinomial regression with variables **Level** and **Male**. Whether the interaction term should be included in the model? [5 pts]

   (b) Based on the model in part (a), calculate the fitted counts for each combination of levels of **Distraction**, **Level** and **Male** (Hint: follow the gator food example) [5 pts]

4. Use data2.dat and treat **Distraction** as an ordinal response variable and other variables as predictors.

   (a) Fit a cumulative logit model with variables **Level** and **Male**. Whether the interaction term should be included in the model? (Hint: you may use the function vglm with logit link) [5 pts]

   (b) Based on the model in part (a), what are $P(\text{Distraction} \leq 0)$ for a male using level 1 and 2? [5 pts]

5. Use data3.dat. Recall that **Distraction** and **LD** are response variables.

   (a) Fit a log linear model including the main effects of **Distraction** and **LD** and the interaction terms between **Level**, **Male**, **Night**, and **Wet**. Could we drop the interaction term? [5 pts]

   (b) Start with the model in Part (a) and carry out a forward selection. [5 pts]

   (c) Draw the association graph for the model selected in Part (b). [5 pts]

   (d) Test whether **Distraction** and **LD** are conditionally independent given other variables in the model from Part (b)? [5 pts]

   (e) Whether the model from Part (b) is equivalent to the main effects logit model or the model with two-way interactions if treating LD as the binary response variable? Justify your answer [5 pts]

   (f) Is the model from Part (b) lack of fit? Justify your answer. [5 pts]

6. Write a brief report to summarize your findings.