

## Psych 6136: Project 2

### Instructions

Several research problems, involving categorical data methods--- multi-way tables, loglinear models, logistic regression or correspondence analysis, count data GLMs are described below.

For TWO of these problems,

- Carry out appropriate analyses, guided by (but not limited to) the questions or suggestions posed. Feel free to make up your own questions. The questions are designed to help you frame a story to tell about the dataset.
- Create meaningful and useful displays to explore the data and explain the results
- Write up a *brief* research report including (a) a problem description, (b) methods of analysis and (c) results section, and (d) a summary/discussion/conclusions section. If you wish, you can include some of the details from your analyses in the previous steps as an appendix, “Supplementary materials”. It would greatly help if you somehow keyed or related your main report with reference to the supplementary materials.

You can use any statistical or other software you like, though you may find that some of the steps or questions are easier to do in R. [You can use `write.csv()` to export an R data set in a form you can import into other software.]

You should submit your report by email (PDF or MS Word); hardcopy submitted to my BSB mailbox would be helpful, but is not required. If you used R, please also submit the .R or .Rmd script(s) you used for your analyses by email. Please name these files along the lines of ‘**YourName-Project2**’. Please practice file name hygiene!

### Notes on presentation:

- Please **don’t use** APA format --- I prefer to read a research report in a format similar to what might appear in a journal article: **single spacing is OK** (actually preferred), with figures and tables presented inline, rather than all at the end. Just leave some marginal space for comments.
- Figures and tables in the report should be numbered, referred to in the text, and have reasonably descriptive captions. (In R, you can use the `fig.cap = “my caption”` chunk option.)
- In your .R or .Rmd files you can set global options for chunks to control figure sizes and other features. See: <https://yihui.org/knitr/options/> for details. It may also be useful to suppress warnings and messages. In an .R script these are included with a special `#+` comment, e.g.,

```
#+ echo=FALSE
knitr::opts_chunk$set(
  warning = FALSE, # avoid warnings and messages in the output
```

```
message = FALSE,  
fig.height = 5, fig.width = 7  
)
```

## Problems

- Dayton Survey data:** The data set `DaytonSurvey` in `vcdExtra` gives a  $2 \times 2 \times 2 \times 2 \times 2$  table in the form of a frequency data frame containing results of a survey of high school students regarding whether they had ever used alcohol (A), cigarettes (C) or marijuana (M), classified by sex (S) and race (R). The goal here is first to understand the associations among the variables A, C, M, and then determine whether and how they differ with sex and race.
  - For the first goal, you can be guided by the questions posed in Exercises 9.1 and 9.2 DDAR. These ignore (collapse over) race and sex, and ask you to fit and plot various loglinear models.
  - For the second goal, extend the analysis to include the variables sex and race as potential explanatory variables of substance use. There are several ways to go about this: You could fit logit models predicting the use of each substance from the remaining variables, or loglinear models that include the `sex*race` association.
  - Use analysis and plots of log odds ratios for the association between A and M in relation to the remaining variables.
- Housing data:** The data set `housing` in the `MASS` package gives a  $3 \times 3 \times 4 \times 2$  table relating satisfaction (Sat) of residents in Copenhagen with their housing to their perceived degree of influence (Infl) on management of the property, the Type of apartment and the degree of contact (Cont) residents have with other residents. Question: How does satisfaction vary with these factors? Exercise 8.2 in DDAR describes some analyses and steps to investigate this question.
- Cormorants data:** This problem is described in DDAR Exercise 11.4. It concerns the counts of male cormorants showing “advertising behavior” to attract females for breeding, observed two or three times a week at six stations in a tree-nesting colony for an entire breeding season. The number of advertising birds was counted and these observations were classified by characteristics of the trees and nests. The goal was to determine how this behavior varies temporally over the season and spatially over observation stations, as well as with characteristics of nesting sites. The response variable is count and other predictors are shown below. See `help(Cormorants, package="vcdExtra")` for further details.
- Survival on the Titanic:** The dataset `vcdExtra::Titanicp` gives individual case data for 1309 passengers on the RMS Titanic. Members of the crew are not included. The goal is to use the methods of the course to understand the binary response `survived`. The available predictors are passenger class (`pclass`), sex, age, number of siblings or spouses aboard (`sibsp`) and number of parents or children aboard (`parch`). Note that age is missing for 263 passengers.
  - Make some incisive exploratory plots showing how survival depends on these predictors and some of their combinations. For example, it might be useful to plot the relation of age to survival broken down by `pclass`, sex, etc. with some suitable smooths, or discrete plots of survival against `sibsp` and/or `parch`.

- b. Ignoring the cases with missing age, fit a main effects model predicting survival from pclass, sex, age, sibsp and parch, all with linear effects. Evaluate the model fit and model terms statistically and with plots of either predicted probabilities or log odds. What do you conclude about this model?
- c. Now develop a more complex model that allows for some degree of non-linearity in the effect of age (for example `poly(age, df)` or `splines::ns(age,df)`), as well as possible interactions of age with sex, pclass, sibsp and parch.

You will find that the distributions of sibsp and parch are highly skewed, so it will be useful to recode both of these to the values [0, 1, 2+]. The following code may be helpful:

```
library(dplyr)
Titanicp <- Titanicp |>
  mutate(sibspF = case_match(sibsp,
                             0 ~ "0",
                             1 ~ "1",
                             2:max(sibsp) ~ "2+")) |>
  mutate(sibspF = ordered(sibspF)) |>
  mutate(parchF = case_match(parch,
                             0 ~ "0",
                             1 ~ "1",
                             2:max(parch) ~ "2+")) |>
  mutate(parchF = ordered(parchF))

# before
table(Titanicp$sibsp, Titanicp$parch)
# after
table(Titanicp$sibspF, Titanicp$parchF)
```

As in part (b), evaluate the model statistically and with some plots of predicted values.