Homework 1, Generalized linear models

STA442 Methods of Applied Statistics

Due 24 September 2019

1 Flies

This is Ex. 6.81 from Faraway (2005). One hundred twenty-five fruit flies were divided randomly into five groups of 25 each. The response was the lifetime of the fruit fly in days. One group was kept solitary, while another was kept individually with a virgin female each day. Another group was given eight virgin females per day. As an additional control the fourth and fifth groups were kept with one or eight pregnant females per day (pregnant fruit flies will not mate). The thorax length of each male was measured as this was known to affect lifetime. The data is fruit fly in the library `faraway`. A complete reference to the data is given in the help file for the dataset.

```
data('fruitfly', package='faraway')
summary(fruitfly)
```

```
##      thorax         longevity         activity
##  Min.   :0.6400   Min.   :16.00   isolated:25
##  1st Qu.:0.7600   1st Qu.:46.00   one     :25
##  Median :0.8400   Median :58.00   low     :25
##  Mean   :0.8224   Mean   :57.62   many    :24
##  3rd Qu.:0.8800   3rd Qu.:70.00   high    :25
##  Max.   :0.9400   Max.   :97.00
```

Use a Gamma generalized linear model to model the lifetimes as a function of the thorax length and activity. Write a brief report (a half to one page of writing) summarizing the problem and the model used, and interpreting the coefficients in your model in terms of their effect on expected lifetime. Write a one-paragraph, non-technical, summary of the results, that might appear in a "Research News" media article about the laboratory in question.

Hints

- consider centering and rescaling variables
- don't show R code in your answer but putting your code in an appendix might help the marker
- format tables and figures nicely
- The code below does not fit a useful model, but it might help you get started

```
glm(thorax ~ longevity + activity, family=Gamma(), data=fruitfly)
```

2 Smoking

Over the course of the next 13 weeks you will be using the 2014 American National Youth Tobacco Survey to become an expert in all matters pertaining to the use of cigars, hookahs, and chewing tobacco amongst

American school children. MS Access and SAS versions of the survey data are available from the Survey's web page. On the pbrown.ca/appliedstats/astwo/data page there is an R version of the 2014 dataset `smoke.RData`, a pdf documentation file `2014-Codebook.pdf`, and the code used to create the R version of the data `smokingData.R`.

The research hypotheses to be investigated using this survey are as follows.

1. Regular use of chewing tobacco, snuff or dip is no more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans, once one accounts for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon.
2. The likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of the different sexes, provided their age, ethnicity, and other demographic characteristics are similar.

Write a short consulting report addressing these hypotheses. This should include the following:

- a one-paragraph summary stating your conclusions, which could be understood by a child health and welfare professional or an executive in the marketing department of a large tobacco firm;
- a writeup of roughly one page of text (not including figures and tables) containing
  - an introduction restating the problem as you've interpreted it in relation to this dataset,
  - a methods section giving the statistical models used (in mathematical notation, not R syntax) and justifying their use, and
  - a results section where the results are described and interpreted; and
- an appendix containing your code.

The report will be assessed in terms of:

- clarity of presentation,
- the use of an appropriate model and implementing it correctly,
- demonstration of an understanding of the statistical models used, and
- drawing conclusions which are consistent with the analysis.

The data

You can obtain the data with:

```
dataDir = "../data"
smokeFile = file.path(dataDir, "smokeDownload.RData")
if (!file.exists(smokeFile)) {
    download.file("http://pbrown.ca/teaching/appliedstats/data/smoke.RData",
        smokeFile)
}
(load(smokeFile))
```

```
## [1] "smoke"        "smokeFormats"
```

The `smoke` object is a `data.frame` containing the data, the `smokeFormats` gives some explanation of the variables. The `colName` and `label` columns of `smokeFormats` contain variable names in `smoke` and descriptions respectively.

- `chewing_tobacco_snuff_or`: RECODE: Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days
- `ever_tobacco_hookah_or_wa`: RECODE: Ever smoked tobacco out of a hookah or waterpipe

The data produced by `smokingData.R` has changed the data in a few ways.

- `RuralUrban` is a flag denoting whether the school the respondent attended was rural or urban.

- `Race` is an R factor recoded from `RaceEth_no_mult_grp`.
- ages have been converted to years from the original categorical variables described in the pdf file

Some words of advice

- Write in sentences and paragraphs.
- Provide captions for ALL figures and tables
- Don't use default axis labels on plots and ensure text on plots is large enough to read comfortably
- Round numbers to 2 or 3 decimal places so tables look tidy.
- Don't show raw R output. Put things in Latex or Markdown tables (using `knitr::kable` or `Hmisc::latex`)
- Give parameter estimates and confidence intervals on the 'natural' scale where possible (probabilities or odds rather than log-odds ratios)

Hints

get rid of 9 year olds because their data is suspicious

```
smokeSub = smoke[smoke$Age >= 10, ]
```

fit a model incapable of answering the research question

```
glm(ever_tobacco_pipe_not_hoo ~ RuralUrban + Race + Age,
  family=binomial, data=smokeSub)
```

```
##
## Call:  glm(formula = ever_tobacco_pipe_not_hoo ~ RuralUrban + Race +
##     Age, family = binomial, data = smokeSub)
##
## Coefficients:
##     (Intercept)  RuralUrbanRural        Raceblack     Racehispanic
##         -8.72748          0.20722         -1.23664         -0.15502
##       Raceasian       Racenative      Racepacific              Age
##         -0.97685         -0.04943         -0.51884          0.35989
##
## Degrees of Freedom: 20027 Total (i.e. Null);  20020 Residual
##   (1939 observations deleted due to missingness)
## Null Deviance:       5884
## Residual Deviance: 5468  AIC: 5484
```

Looks like white kids smoke pipes more than anyone else.

References

Faraway, J.J. (2005). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. URL: http://www.tandfebooks.com/isbn/9780203492284.