Homework 1, Generalized linear models

Methods of Applied Statistics II

Due 29 Jan 2016

1 Short answer (20 marks)

1.1 Simulation study (5)

Use the code below to simulate 100 datasets (with different random seeds, naturally) and fit the 'correct' model to it.

```
set.seed(0)
x = seq(-10, 10, len=40)
off = rep(c(1,-1), c(25, length(x)-25))
y = rpois(length(x), exp(off + 0.5 +0.2*x))
summary(glm(y~x+offset(off), family='poisson'))[['coefficients']]
```

```
##              Estimate Std. Error  z value     Pr(>|z|)
## (Intercept) 0.5810477 0.09868293 5.888026 3.908359e-09
## x           0.1871256 0.01888513 9.908621 3.818818e-23
```

1. Compute the coverage probability of a 2 standard error confidence interval for the coefficient on x and assess whether this is a useful way to construct 95% confidence intervals.
2. Assess whether the $\hat{\beta}$ coefficient for x is well approximated by a Normal distribution centred on 0.2.
3. Calculate 100 likelihood ratio statistics for testing $\beta = 0.2$. Does this appear roughly chi-squared with the appropriate degrees of freedom?

1.2 Distribution functions (5)

This question concerns each of the five distribution functions below.

- Zero-inflated Poisson;

- Gamma;
- Weibull;
- log-Normal; and
- Negative binomial.

1. Derive parameters for each distribution which will result in random variables with mean 2 and variance 3. Derive the parameters analytically (to the extent possible) and show your work, and give final numeric values using R to evaluate special functions or solve equations numerically when required.
2. Produce a figure with all five density functions on the same graph.

3. Find the 99% upper quantile for each distribution.
4. Simulate 20 realisations from each distribution and compute their sample mean and sample variance.

You may not use the `dlnorm` family of functions listed here stat.ethz.ch/R-manual/R-devel/library/stats/html/Lognormal.html or the `ZIP` functions from www.inside-r.org/packages/cran/gaml Use write your own log-normal and zero-inflated density and simulation code using `dnorm` and `dpois`.

1.3 Data analysis (5)

This is Ex. 6.81 from Faraway (2006). 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`.

```
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. Interpret the coefficients in your model, in terms of their effect on

expected lifetime, and assess the fit of your model with appropriate diagnostics. Give 95% confidence intervals for any contrasts you think may be relevant. A complete reference to the data is given in the help file for the dataset. Write a one-paragraph, non-technical, summary of the results, that might appear in a "Research News" media article about the laboratory in question.

1.4 Discussion (5)

After completing part 2's tobacco assignment, read one or both of Breiman (2001) and Donoho (2015). Write a short discussion (roughly 1 Latex page) about Brieman's two cultures in the context of your tobacco analysis. State some of the criticisms that Brieman or Cox might make about your report, and either justify your approach or explain how and why you would do the analysis differently. You may wish to address the research hypotheses and the secondary problem differently.

2 Report (20 marks)

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/teaching/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, stuff 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 same sex, provided their age, ethnicity, and other demographic characteristics are similar.

A secondary problem, which is not a formal research hypothesis, is to quantify how the use of chewing tobacco changes with age, sex, and ethnic group.

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

- a summary of a couple of paragraphs 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 detailed writeup of roughly two pages 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/astwo/data/smoke.RData',
            smokeFile
            )
}
print(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.

```
for(D in c('chewing_tobacco_snuff_or', 'ever_tobacco_hookah_or_wa')){
  cat("- `", D,
  "`:  ",
  as.character(smokeFormats[match(D, smokeFormats[,'colName']), 'label']),
  '\n\n', sep='')
}
```

- `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)

References

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statist. Sci.* 16 (3). The Institute of Mathematical Statistics: 199–231. doi:10.1214/ss/1009213726.

Donoho, David. 2015. "50 Years of Data Science." https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf.

Faraway, JJ. 2006. "Extending the Linear Model with R. Chapman Hall." *CRC, Boca Raton*.