Homework 1, Generalized linear models

STA442 Methods of Applied Statistics

Due 9 Oct 2020

1 Affairs

The 'Affairs' dataset in the `AER` package is described at

vincentarelbundock.github.io/Rdatasets/doc/AER/Affairs.html.

It contains data collected from 600 married readers of the American magazines *Redbook* and *Psychology Today* in 1969, who gave information about the frequency they had extramarital sex.

The research question of interest concerns the effect of having children on the chances that men and women have affairs. It is hypothesized that becoming a mother will make a woman less likely to have extramarital sex, whereas men are believed to be more likely to have affairs once they become fathers. The individual making this hypothesis has a knowledge of human behaviour based entirely on viewing TV sitcoms, and believes women become insular and overly protective once they become mothers but new fathers are likely to feel neglected and constrained.

Use a logistic regression model to explore the factors influencing the chance that an individual has an affair using the `Affairs` dataset. Age, the numer of years married, and religiousness of the individuals concerned are important confounders which must be taken into account.

1. 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 having affairs.
2. Write a one-paragraph, non-technical, summary of the results, that might appear in a "Research News" media article about the the scientists carrying out this study.

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

```
data('Affairs', package='AER')
Affairs$ever = Affairs$affair > 0
Affairs$religious = factor(Affairs$religiousness,
  levels = c(2,1,3,4,5), labels = c('no','anti','low','med','high'))
summary(glm(
  ever ~ gender:children + age + yearsmarried + religious,
  data=Affairs, family='binomial'))$coef
```

```
##                              Estimate Std. Error    z value    Pr(>|z|)
## (Intercept)              -0.42331185 0.51440987 -0.8229077 0.410560471
## age                      -0.03634535 0.01790000 -2.0304663 0.042309165
## yearsmarried              0.10495138 0.03203713  3.2759297 0.001053148
## religiousanti             0.70564305 0.35612254  1.9814613 0.047539560
## religiouslow              0.27522689 0.26880682  1.0238836 0.305890265
```

```
## religiousmed             -0.72788170 0.27430012 -2.6535960 0.007963912
## religioushigh            -0.66388456 0.37097076 -1.7895873 0.073520279
## genderfemale:childrenno  -0.91706033 0.38295256 -2.3947100 0.016633514
## gendermale:childrenno    -0.25258442 0.36056885 -0.7005165 0.483604832
## genderfemale:childrenyes -0.26387204 0.22859361 -1.1543281 0.248365665
```

2 Smoking

Over the course of the next 13 weeks you will be using the 2019 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 2019 dataset `smoke.RData`, a pdf documentation file `2019-nyts-codebook-p.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. Smoking of cigars, cigarillos or little cigars 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 cigar smoking is a rural phenomenon.
2. The likelihood of having used a electronic cigarettes 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.

- `ever_chewing_tobacco_snu`: Have you ever used chewing tobacco, snuff, or dip, such as Copenhagen, Grizzly, Skoal, or Longhorn, even just a small amount?
- `Age_first_hookah_or_water`: How old were you when you first tried smoking tobacco in a hookah or waterpipe, even one or two puffs?
- `ever_cigars_cigarillos_or`: Have you ever tried smoking cigars, cigarillos, or little cigars, such as Swisher Sweets, Black and Mild, Garcia y Vega, Cheyenne, White Owl, or Dutch Masters, even one or two puffs?
- `ever_ecigarette`: Have you ever used an e-cigarette, even once or twice?

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[which(smoke$Age >= 10), ]
```

fit a model incapable of answering the research question

```
summary(glm(ever_cigars_cigarillos_or ~ 0+ RuralUrban+Race+Sex + Age,
  family=binomial, data=smokeSub))$coef
```

```
##                    Estimate Std. Error    z value      Pr(>|z|)
## RuralUrbanUrban -7.54392137 0.19236710 -39.216277  0.000000e+00
## RuralUrbanRural -7.14244510 0.19014639 -37.562874  0.000000e+00
## Raceblack        0.42862587 0.06395822   6.701654  2.060739e-11
## Racehispanic    -0.05658227 0.05307857  -1.066010  2.864193e-01
## Raceasian       -1.26215116 0.17160260  -7.355082  1.908103e-13
## Racenative       0.28275344 0.20528600   1.377363  1.683999e-01
## Racepacific      0.43177279 0.28018109   1.541049  1.233049e-01
## SexF            -0.38097546 0.04580092  -8.318075  8.940342e-17
## Age              0.37384576 0.01193497  31.323571 2.229010e-215
```