CRSP 500 / EPBI 500

Design and Analysis of Observational Studies

Old Homework 1 used in 2016 and earlier

Professor Thomas E. Love

Version: This document published **January 6, 2018**.

1 General Instructions

This assignment required students to analyze some data, and prepare a report. That report needed to be in the form of a Word or PDF file, along with a separate R script or (much better) R Markdown file which allows me to completely replicate the analysis. Note that the answer sketch for this document includes both the Markdown and PDF versions.

1. The files you develop should be `YOURNAME-500hw1.docx` or `YOURNAME-500hw1.pdf` and `YOURNAME-500hw1.Rmd` or `YOURNAME-500hw1.R`, please.
2. Do professional work with this little problem. What do I mean by this?
    - Properly labeled graphs/figures are a *minimum* expectation for graduate school.
    - Use complete English sentences to describe your findings.
    - Make sure that the answers include enough of the question that your text responses (in addition to the graphs) stand on their own. Be sure to address all three tasks.
    - Present edited code, making an effort to delete false starts, and comment liberally.
    - Use words I know, without simply repeating my explanations verbatim, please.
3. You are welcome to discuss Assignment 1 with anyone, including myself, or your colleagues, but your answer must be prepared by you alone.
4. If you are confused by the assignment, or stuck in the development of your response, please ask questions!

2 Data

The **hw1.csv** data file is available at the Assignments page of the course web site.

- The file includes 135 subjects, the first 40 of whom have received a particular `treatment` and the remaining 95 of whom have not received it.
- Also provided are five meaningful predictors of treatment status, labeled (imaginatively) `cov1`, `cov2`, `cov3`, `cov4` and `female`.
- Covariates 1-4 are continuous covariates, gathered at varying levels of precision. The `female` variable indicates gender (1 = female, 0 = male.)
- Happily, there are no missing values in the data.

3 Tasks

1. Build a logistic regression model using the main effects of the five predictors to predict `treatment` status.
   - Use R to add two columns to the data set, specifically the fitted probability (according to your logistic regression model) of being treated, and the linear component of the logistic regression model (i.e. the logit of the probability of being treated.)

   As a hint, partial R code you might use to do this work follows...

```
m1 <- glm(treatment=="Treated" ~ cov1 + cov2 + cov3 + cov4 + female,
                  family=binomial(), data=hw1)
hw1$linpred <- m1$linear.predictors
hw1$prob <- m1$fitted.values
View(hw1) # shows newly added columns in spreadsheet (capitalize V)
```

2. Next, summarize the resulting probabilities across the untreated and treated patients in an appropriate and attractive manner.
   - Raw R code is rarely attractive on its face - build something brief, effective and appropriate for a presentation.
   - Of course, we'd expect that the average probability of being treated will be higher in the patients who are actually treated. Verify that this is the case, in a short **numerical and graphical** summary of your findings.
3. How much overlap is there between the fitted probabilities of the treated patients and the fitted probabilities of the untreated patients?
   - A **graph** of this overlap (perhaps a boxplot, but a better option would be a dot chart or density plot of some sort; creativity is welcome here) is crucial, supplemented by a short written description of your findings.