IPS9 in R: Logistic Regression (Chapter 14)

Nicholas Horton (nhorton@amherst.edu)

January 19, 2019

Introduction and background

These documents are intended to help describe how to undertake analyses introduced as examples in the Ninth Edition of *Introduction to the Practice of Statistics* (2017) by Moore, McCabe, and Craig.

More information about the book can be found here. The data used in these documents can be found under Data Sets in the Student Site. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at https://nhorton.people.amherst.edu/ips9/.

This work leverages initiatives undertaken by Project MOSAIC (http://www.mosaic-web.org), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the `mosaic` package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignettes (http://cran.r-project.org/web/packages/mosaic). A paper describing the mosaic approach was published in the *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Chapter 14: Logistic Regression

This file replicates the analyses from Chapter 14: Logistic regression.

First, load the packages that will be needed for this document:

```
library(mosaic)
library(readr)
```

Section 14.1: The Logistic Regression Model

Example 14.3: Comparing the proportions of female and make Instagram users

```
Instagram <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter14/EG14-03INSTAGR.csv")
```

```
## Parsed with column specification:
## cols(
##   Sex = col_character(),
##   SexNum = col_double(),
##   User = col_character(),
##   Count = col_double()
## )
```

```
Instagram
```

```
## # A tibble: 4 x 4
##   Sex    SexNum User  Count
##   <chr>   <dbl> <chr> <dbl>
## 1 1Women      1 Yes     328
## 2 1Women      1 No      209
## 3 2Men        0 Yes     234
## 4 2Men        0 No      298
```

```r
InstaMatrix <- matrix(c(Instagram$Count), nrow = 2)
rownames(InstaMatrix) <- c("Yes", "No")
colnames(InstaMatrix) <- c("Women", "Men")
InstaMatrix
```

```
##     Women Men
## Yes   328 234
## No    209 298
```

```r
oddsRatio(InstaMatrix, verbose = TRUE)
```

```
##
## Odds Ratio
##
## Proportions
##     Prop. 1:  0.5836
##     Prop. 2:  0.4122
##   Rel. Risk:  0.7063
##
## Odds
##      Odds 1:  1.402
##      Odds 2:  0.7013
##  Odds Ratio:  0.5003
##
## 95 percent confidence interval:
##   0.6232 < RR < 0.8005
##   0.3921 < OR < 0.6384
## NULL
```

```
## [1] 0.5003478
```

Example 14.6: Is a movie going to be profitable?

```r
Movies <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter14/EG14-06MOVIES.csv")
```

```
## Parsed with column specification:
## cols(
##   Title = col_character(),
##   Budget = col_double(),
##   USRevenue = col_double(),
##   Opening = col_double(),
##   Theaters = col_double(),
##   Opinion = col_double(),
##   LOpening = col_double(),
##   Profit = col_double(),
##   Profita = col_character()
## )
```

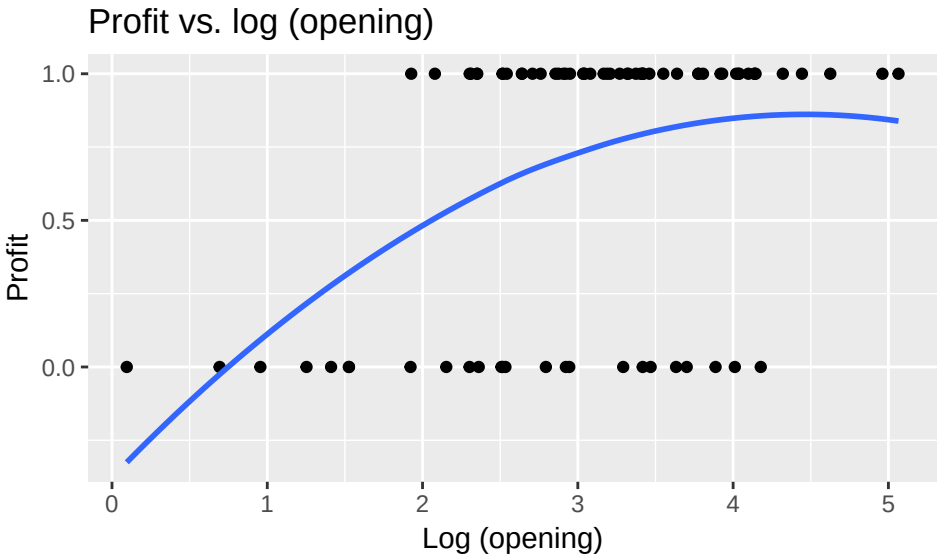
```r
# Log odds
moviemod <- glm(as.factor(Profit) ~ LOpening, data = Movies, family = "binomial")
moviemod
```

```
##
## Call:  glm(formula = as.factor(Profit) ~ LOpening, family = "binomial",
##     data = Movies)
##
```

```
## Coefficients:
## (Intercept)     LOpening
##      -2.556        1.125
##
## Degrees of Freedom: 77 Total (i.e. Null);  76 Residual
## Null Deviance:       97.85
## Residual Deviance: 83.04     AIC: 87.04
```

```
# Figure 14.3, page 8
gf_point(Profit ~ LOpening, data = Movies) %>%
  gf_smooth(span = 2) %>%
  gf_labs(x = "Log (opening)", title = "Profit vs. log (opening)") # to adjust smoothness
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Section 14.2: Inference for Logistic Regression

```
msummary(moviemod)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.556      1.009  -2.532 0.011331 *
## LOpening       1.125      0.339   3.320 0.000901 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 97.852  on 77  degrees of freedom
## Residual deviance: 83.035  on 76  degrees of freedom
## AIC: 87.035
##
## Number of Fisher Scoring iterations: 4
```

Example 14.7: Software output

```
Instagram <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter14/EG14-07INSTAGR.csv")
```

```
## Parsed with column specification:
## cols(
##   Sex = col_character(),
##   SexNum = col_double(),
##   User = col_character(),
##   Count = col_double()
## )
```

```
# XX not sure how to do this
```

Example 14.8: An insecticide for aphids

```
Insecticide <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter14/EG14-08INSECTS.csv")
```

```
## Parsed with column specification:
## cols(
##   Lconc = col_double(),
##   Kill = col_character(),
##   KillNumeric = col_double(),
##   NUMBER = col_double()
## )
```

```
# Figure 14.8, page 12
#insectmod <- glm()

#gf_point()
```