Chapter 1 Exercises
Applied Logistic Regression, Hosmer

Héctor Lira Talancón

July 2019

1 Setup

The datasets used in these exercises can be found in this link:

https://wiley.mpstechnologies.com/wiley/BOBContent/searchLPBobContent.do

Input the following information to find the datasets related to this textbook:
- ISBN: 9780470582473
- Title: Applied Logistic Regression
- Author/Editor: Stanley Lemeshow , David W Hosmer , Rodney X Sturdivant

2 Exercises

1. Dataset used: ICU dataset

   (a) Let Y be our response variable, STA, and x be our independent variable, AGE. Then, the logistic regression model of STA on AGE is stated as:

   $$E[Y|x] = \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

   The logit transformation of our response variable is stated as:

   $$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

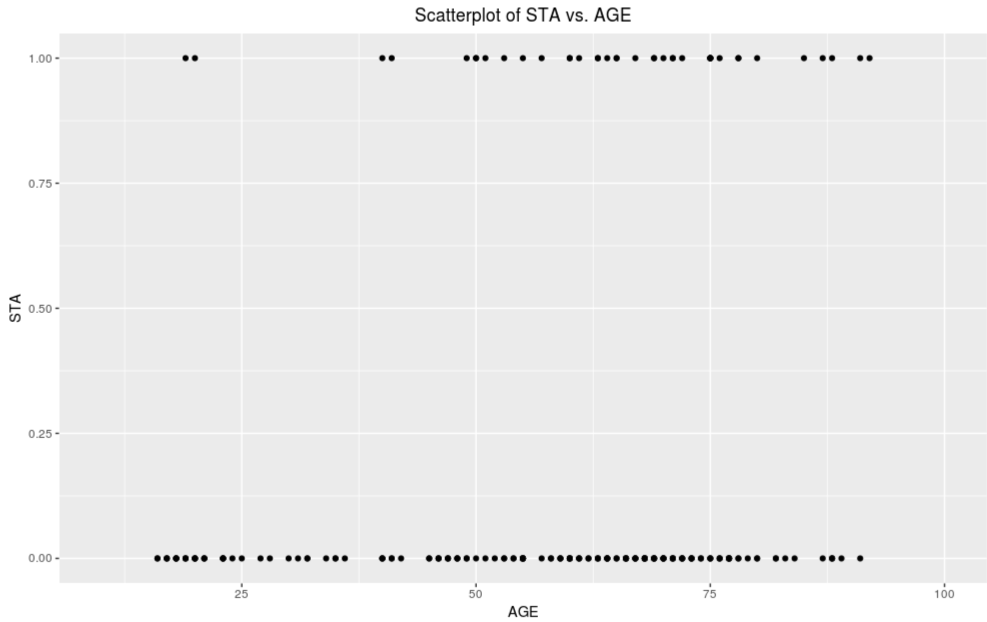
   Given that our response variable, STA, is dichotomous, it is preferred that we use a logistic model over a linear model.

   (b) Scatterplot of STA vs. AGE:

```
library(ggplot2)
library(data.table)

icu_data <- fread("datasets/ICU/ICU.txt", header = T)

ggplot(data = icu_data) +
  geom_point(aes(y = STA, x = AGE)) +
  xlim(c(10,100)) +
  ylim(c(0,1)) +
  ggtitle("Scatterplot of STA vs. AGE") +
  theme(plot.title = element_text(hjust = 0.5))
```



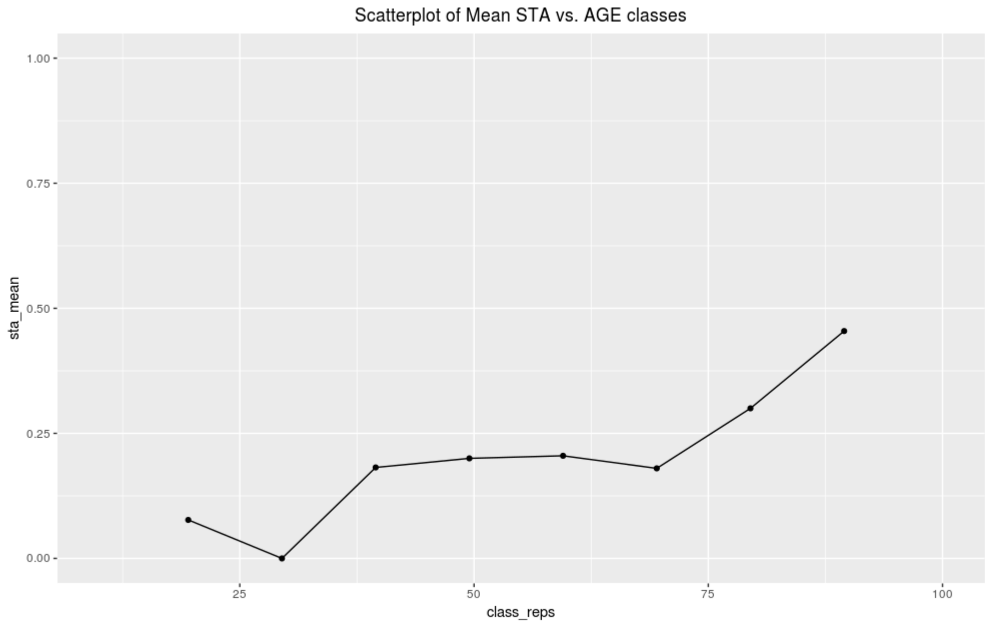
(c) Taking the AGE intervals $[15, 25)$, $[25, 35)$, $[35, 45)$, $[45, 55)$, $[55, 65)$, $[65, 75)$, $[75, 85)$, $[85, 95]$, we plot the mean STA for each interval:

```
library(dplyr)

intervals <- 15 + 10 * 0:8
class_rep <- rowMeans(cbind(head(intervals, -1),
                            intervals[-1] -1))

icu_data_summary <- icu_data %>%
  mutate(age_intervals = cut(AGE, breaks = intervals,
                             include.lowest = T,
                             right = F)) %>%
  group_by(age_intervals) %>%
  summarise(sta_mean = mean(STA)) %>%
  mutate(class_reps = class_rep)

ggplot(data = icu_data_summary,
       aes(y = sta_mean, x = class_reps)) +
  geom_line() +
  geom_point() +
  xlim(c(10,100)) +
  ylim(c(0,1)) +
  ggtitle("Scatterplot of Mean STA vs. AGE classes") +
  theme(plot.title = element_text(hjust = 0.5))
```



(d) Let $\beta = (\beta_0, \beta_1)$, let y_i be the i-th observation for the STA variable, and let x_i be the i-th observation of the AGE variable. Then, the likelihood of the logistic regression model is stated as:

$$l(\beta) = \prod_{i=1}^{200} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{(1-y_i)}$$

The log-likelihood expression for our logistic regression model is stated as:

$$L(\beta) = \ln\left[l\left(\beta\right)\right] = \sum_{i=1}^{200} \left\{y_i \pi(x_i) + (1 - y_i)\left[1 - \pi\left(x_i\right)\right]\right\}$$

(e) Fitting a logistic regression model to our data, we obtain the following estimates: