Chapter 1 Exercises

Applied Logistic Regression - DAVID W. HOSMER, JR, et al.

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:

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

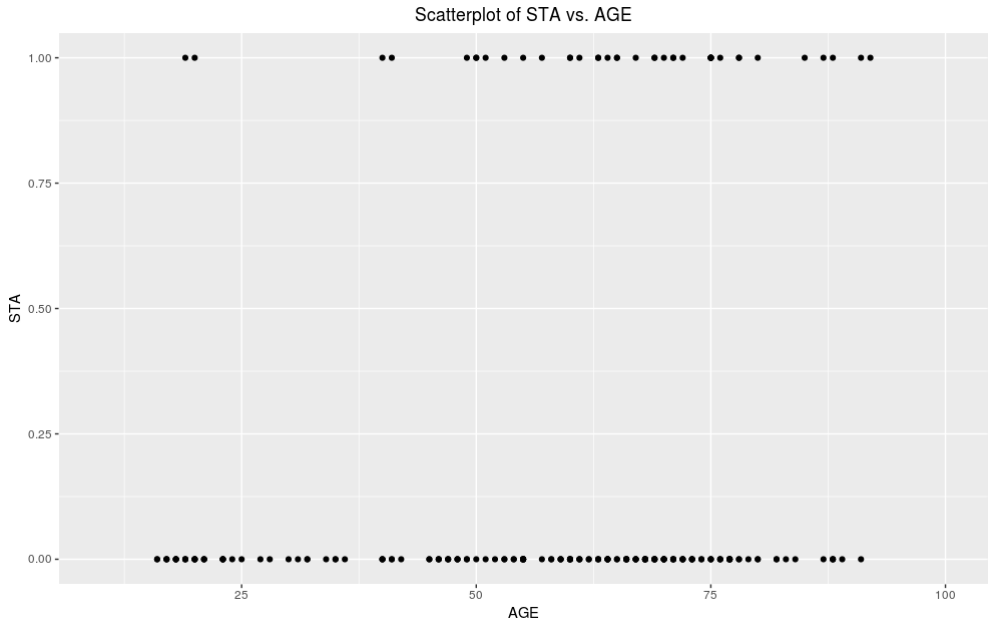
   The logit transformation of our response variable is stated as:

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

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

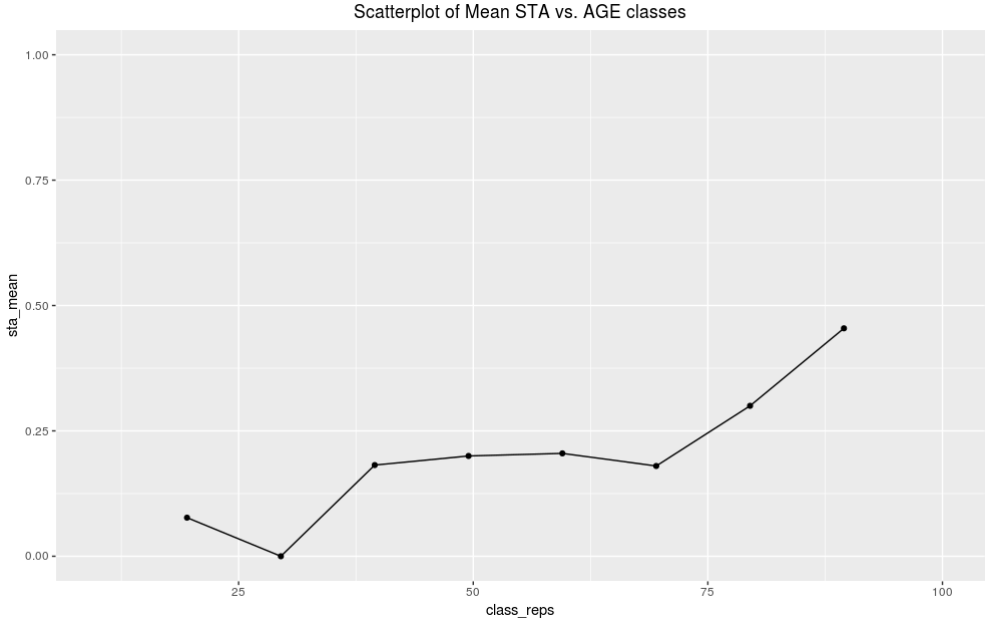
   (b) Scatterplot of STA vs. AGE:

```
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:

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

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

```
library(dplyr)

logistic_model <- glm(STA ~ AGE,
                      data = icu_data,
                      family = binomial)
summary(logistic_model)

Output:
Call:
glm(formula = STA ~ AGE, family = binomial, data = icu_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9536  -0.7391  -0.6145  -0.3905   2.2854

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05851    0.69608  -4.394 1.11e-05 ***
AGE          0.02754    0.01056   2.607  0.00913 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

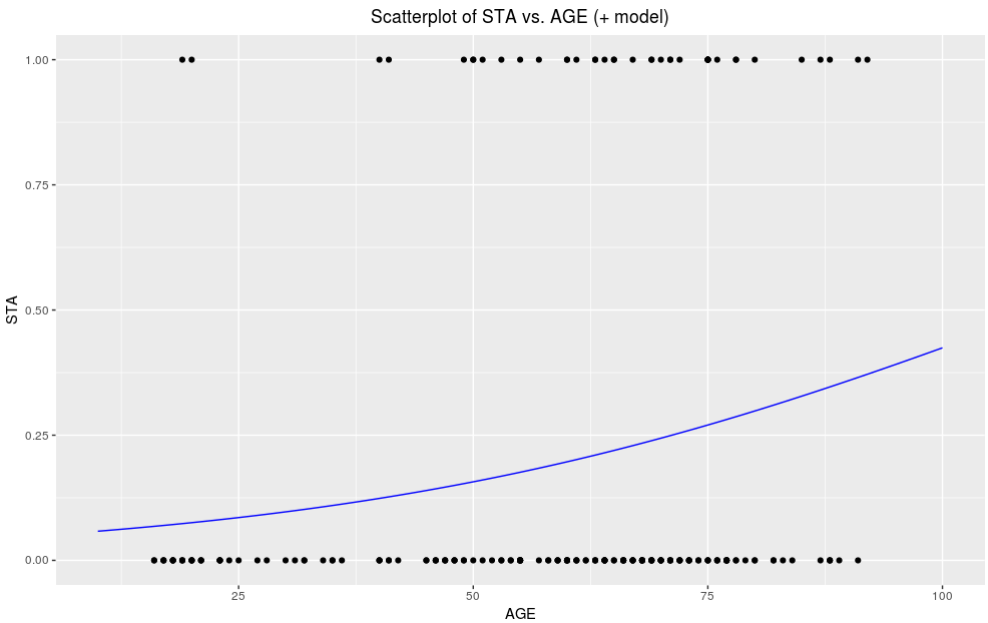
```
    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 192.31  on 198  degrees of freedom
AIC: 196.31

Number of Fisher Scoring iterations: 4
```

Given the above estimates, the equation for our logistic regression model is stated as follows:

$$\hat{\pi}(x) = \frac{e^{-3.0585+0.0275x}}{1+e^{-3.0585+0.0275x}}$$

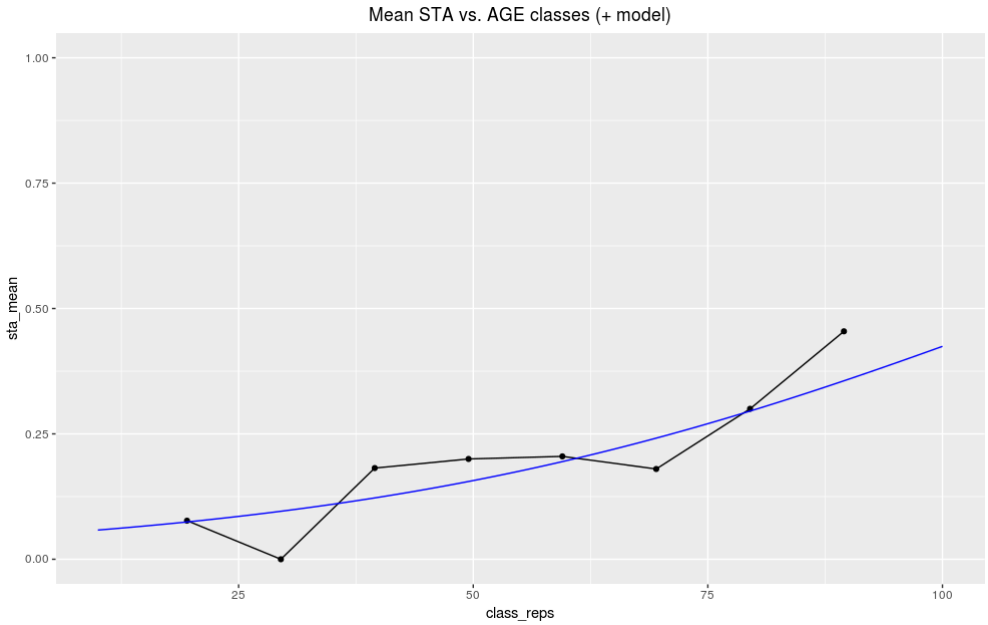
We plot this equation in the scatterplots from (b) and (c). The commented section in the code below displays another way of plotting the logistic regression curve in a more simpler way, without having to define `logistic_model_function`:

```
logistic_model_function <- function(x){
  logit_prob <- logistic_model$coefficients[1] +
    logistic_model$coefficients[2] * x

  return(exp(logit_prob) / (1 + exp(logit_prob)))
}

ggplot(data = icu_data, aes(y = STA, x = AGE)) +
  geom_point() +
  # geom_smooth(method = "glm",
  #             method.args = list(family = "binomial"),
  #             se = F) +
  stat_function(fun = logistic_model_function,
                color = 'blue') +
  xlim(c(10,100)) +
  ylim(c(0,1)) +
  ggtitle("Scatterplot of STA vs. AGE (+ model)") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(data = icu_data_summary,
       aes(y = sta_mean, x = class_reps)) +
  geom_line() +
  geom_point() +
```

```
  stat_function(fun = logistic_model_function,
                color = 'blue') +
  xlim(c(10,100)) +
  ylim(c(0,1)) +
  ggtitle("Mean STA vs. AGE classes (+ model)") +
  theme(plot.title = element_text(hjust = 0.5))
```



(f) We will assess the significance of the variable AGE using the likelihood ratio test, the Wald test, and the Score test.

i. **Likelihood Ratio Test**
To assess the significance of the variable AGE using the likelihood ratio test, we must calculate the following statistic:

$$
\begin{aligned}
G &= -2\ln\left[\frac{\text{likelihood without AGE}}{\text{likelihood with AGE}}\right] \\
&= -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{200}\hat{\pi}_i^{y_i}\left(1-\hat{\pi}_i\right)^{(1-y_i)}}\right] \\
&= 2\left\{\sum_{i=1}^{200}\left[y_i\ln(\hat{\pi}_i)+(1-y_i)\ln(1-\hat{\pi}_i)\right]-\left[n_1\ln(n_1)+n_0\ln(n_0)-n\ln(n)\right]\right\}
\end{aligned}
$$

Where $n_1 = \sum_{i=1}^{200} y_1$ and $n_0 = \sum_{i=1}^{200}(1-y_i)$. Note that $n = n_1 + n_0$. G is known to follow a chi-squared distribution with 1 degree of freedom.

We calculate the statistic by first calculating the likelihood of the model without variables and then the likelihood of the model with the AGE variable. We take the natural logarithm of the ratio of the likelihoods and multiply that by -2:

```
likelihood_without_variables <- function(y){
  n1 <- sum(y == 1)
  n0 <- sum(y == 0)
  n <- n1 + n0

  likelihood <- ((n1/n)^n1) * ((n0/n)^n0)
```

```
  return(likelihood)
}

logistic_likelihood <- function(y, model_predictions){
  likelihood <- prod((model_predictions^y) *
                       (1-model_predictions)^(1-y))

  return(likelihood)
}

predictions <- predict(logistic_model,
                       type = 'response')

G <- -2 * log(likelihood_without_variables(icu_data$STA) /
              logistic_likelihood(icu_data$STA,
                                  predictions))

G

Output:
 [1] 7.854589
```

Running the likelihood ratio test involves calculating:

$$P\left[\chi^2(1) > 7.8546\right] = 1 - P\left[\chi^2(1) \leq 7.8546\right]$$

and comparing this to a significance level of, say, 0.05.

```
1 - pchisq(G, df = 1)

Output:
 [1] 0.005069187
```

Given that $0.0051 < 0.05$, we reject the null hypothesis that $\beta_1 = 0$ using the likelihood ratio test.

ii. **Wald Test**

The Wald test is the one used to display the model summary in R. In the summary table from e), we can see the p-values for each of the model coefficients. However, we will calculate these values manually to show how they work.

The statistic used here is calculated as:

$$W = \frac{\hat{\beta}_i}{\hat{\text{SE}}(\hat{\beta}_i)}$$

Where W follows a standard normal distribution under the null hypothesis. We already know that $\hat{\beta}_1 = 0.0275$. In this stage of the textbook, we have not seen the formula to calculate the standard error of the coefficients. We will use the standard errors calculated for us using R:

```
coef(summary(logistic_model))[2, 2]

Output:
 [1] 0.01056416
```

Then, our statistic has the following value:

```
wald_statistic <- coef(logistic_model)[2] /
                  coef(summary(logistic_model))[2, 2]
wald_statistic
```

```
Output:
 [1] 2.607174
```

The p-value in the Wald test is two-tailed, meaning we must calculate the probability of W being lower and higher than a given threshold. The p-value is calculated as follows (using the symmetrical property of the standard normal distribution):

$$P\left[|z| > 2.6072\right] = 2P[z > 2.6072] = 2\left(1 - P\left[z \leq 2.6072\right]\right)$$

```
2 * (1 - pnorm(wald_statistic))

Output:
 [1] 0.009129303
```

Given that $0.0091 < 0.05$, we reject the null hypothesis that $\beta_1 = 0$ using the Wald test.

iii. **Score Test**

The statistic used in the Score Test is calculated as:

$$\text{ST} = \frac{\sum_{i=1}^{200} x_i \left(y_i - \bar{y}\right)}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^{200} (x_i - \bar{x})^2}}$$

Where ST follows a standard normal distribution under the null hypothesis.

We compute the ST statistic:

```
compute_st_statistic <- function(x, y){
  st_numerator <- sum(x * (y - mean(y)))
  st_denominator <- sqrt(mean(y) *
                           (1 - mean(y)) *
                           sum((x - mean(x))^2))

  st_statistic <- st_numerator / st_denominator

  return(st_statistic)
}

st_statistic <- compute_st_statistic(icu_data$AGE,
                                     icu_data$STA)
st_statistic

Output:
 [1] 2.679339
```

As in the Wald test, we calculate the p-value as:

$$P\left[|z| > 2.6793\right] = 2P[z > 2.6793] = 2\left(1 - P\left[z \leq 2.6793\right]\right)$$

```
2 * (1 - pnorm(st_statistic))

Output:
 [1] 0.007376774
```

Given that $0.0073 < 0.05$, we reject the null hypothesis that $\beta_1 = 0$ using the Score test.

We compute a summary of the results from the three tests:

Test	Statistic	p-value	Interpretation
Likelihood Ratio	7.8546	0.0051	Reject H_0.
Wald	2.6072	0.0091	Reject H_0.
Score	2.6793	0.0074	Reject H_0.

All tests agree on rejecting the null hypothesis ($\beta_1 = 0$).

Note how the statistics from all tests are very similar to one another ($\sqrt{G} = 2.8026$). The assumptions needed in these tests involve having a sufficiently large sample and enough observations where $y = 0$ and $y = 1$.

Finally, we calculate the deviance of the model. R computes the deviance in the summary. We will calculate this value by hand to illustrate how it is calculated.

The deviance, D, of a logistic regression model is calculated as:

$$D = -2\ln(\text{likelihood of the fitted model})$$

```
-2 * log(logistic_likelihood(icu_data$STA, predictions))

Output:
 [1] 192.3064
```

Which matches the deviance calculated by R:

```
logistic_model$deviance

Output:
 [1] 192.3064
```

(g) We will now compute the 95% confidence interval for the coefficient of the variable AGE. The confidence interval is calculated as:

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\hat{\text{SE}}(\hat{\beta}_1)$$

```
alpha <- 0.05
se_b1 <- coef(summary(logistic_model))[2, 2]

logistic_model$coefficients[2] +
  c("lower" = -qnorm(1 - alpha/2) * se_b1,
    "upper" = qnorm(1 - alpha/2) * se_b1)

Output:
      lower       upper
 0.00683723 0.04824799
```

This can also be calculated using an in-built function in R:

```
confint.default(logistic_model,
                parm = 2,
                level = 1 - alpha)

Output:
          2.5 %      97.5 %
AGE 0.00683723 0.04824799
```

The difference between `confint.default` and `confint` is that the latter calculates the profile confidence intervals, while the former calculates the Wald-based confidence intervals.

The results suggest that the change in the log-odds of STA per one unit (year) increase in AGE is 0.0275. This change could be as little as 0.0068 or as much as 0.0482 with 95% confidence.

(h) The covariance matrix of the logistic regression model coefficients is obtained as:

```
summary(logistic_model)$cov.scaled

Output:
             (Intercept)           AGE
(Intercept)  0.484529087 -0.0071029945
AGE         -0.007102994  0.0001116015
```

We will now make the logit and logistic probability prediction for a 60-year old subject. This can be done manually or using in-built R functions. The logit prediction is made using the following formula:

$$\hat{g}(x) = \beta_0 + \beta_1 x$$

Then, the logit prediction is:

```
x <- 60
sum(logistic_model$coefficients * c(1, x))

Output:
 [1] -1.405957
```

Or, using in-built R functions:

```
predict(logistic_model,
        newdata = data.frame(AGE = x),
        type = 'link')

Output:
        1
-1.405957
```

The logistic probability prediction is calculated as:

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$

Then, the logistic probability prediction is:

```
logit_prediction <- sum(logistic_model$coefficients * c(1, x))
exp(logit_prediction) / (1 + exp(logit_prediction))

Output:
 [1] 0.1968726
```

Or, using in-built R functions:

```
predict(logistic_model,
        newdata = data.frame(AGE = x),
        type = 'response')

Output:
        1
0.1968726
```

To calculate the endpoints of the Wald-based 95% confidence interval for the estimated logit, we use the following formula:

$$\hat{g}(x) \pm z_{1-\alpha/2}\hat{\text{SE}}\left[\hat{g}(x)\right]$$

Then, the endpoints of this confidence interval is:

```
alpha <- 0.05

var_g <- summary(logistic_model)$cov.scaled[1,1] +
  (x^2) * summary(logistic_model)$cov.scaled[2,2] +
  2 * x * summary(logistic_model)$cov.scaled[2, 1]

se_g <- sqrt(var_g)

g <- predict(logistic_model,
             newdata = data.frame(AGE = x),
             type = 'link')

g + c(lower = -qnorm(1 - alpha/2) * se_g,
      g = 0,
      upper = qnorm(1 - alpha/2) * se_g)
```

Output:

```
    lower          g      upper
-1.767012  -1.405957  -1.044901
```

To calculate the endpoints of the Wald-based 95% confidence interval for the estimated logistic probability, we use the following formula:

$$\frac{e^{\hat{g}(x)\pm z_{1-\alpha/2}\hat{\text{SE}}[\hat{g}(x)]}}{1+e^{\hat{g}(x)\pm z_{1-\alpha/2}\hat{\text{SE}}[\hat{g}(x)]}}$$

Then, the endpoints of this confidence interval is:

```
g_intervals <- g + c(lower = -qnorm(1 - alpha/2) * se_g,
                     g = 0,
                     upper = qnorm(1 - alpha/2) * se_g)

exp(g_intervals) / (1 + exp(g_intervals))
```

Output:

```
    lower          g      upper
0.1459143  0.1968726  0.2602054
```

The value of 0.1969 is an estimate of the proportion of 60 year-old subjects that survive at hospital discharge. This proportion could be as little as 0.1459 or as high as 0.2602 with 95% confidence.

2. Dataset used: Myopia Study