Introduction to Logistic Regression Part II

Dr. Kiah Wah Ong

The Logit Model

In Part I, we see that if we let

$$\pi = \Pr(Y = 1|X_1 = x_1, \cdots, X_p = x_p)$$
$$= \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

then

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}.$$

Taking the natural logarithm leads to

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

We now want to see how to approximate the parameters $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$.

Estimation in Logistic Regression

There is no closed-form solution in finding $\hat{\beta}_i$ in logistic regression.

This is different from the situation we encountered in linear regression.

Also, unlike the least squares method in linear regression, we need to employ the maximum likelihood method in estimating the regression parameters.

Let us look at how we deal with this new situation by looking at the case when we have a simple logistic regression.

Estimation in Simple Logistic Regression

Let us consider a simple logistic regression, with data taken to be $(x_i, y_i),\ i = 1, \cdots, n$, where y_i be the result (either 1 if success or 0 if failure).

The parameters β_0 and β_1 of the logistic regression function below are assumed to be unknown and need to be estimated.

$$\ln\left(\frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)}\right) = \beta_0 + \beta_1 X$$

Estimation in Simple Logistic Regression

Let us denote

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

Using the Bernoulli density function, we write

$$\begin{aligned}\Pr(Y_i = y_i) &= p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i} \\ &= \left(\frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}\right)^{y_i}\left(\frac{1}{1 + e^{\beta_0+\beta_1 x_i}}\right)^{1-y_i}\end{aligned}$$

For $y_i = 0, 1$

Estimation in Simple Logistic Regression

In order to use the maximum likelihood method. We would like to calculate the total probability of observing all of the data $(x_i, y_i),\ i = 1, \cdots, n$, namely,

$$\Pr(Y_i = y_i, i = 1, \cdots, n)$$

This can get very complicated, hence we need the assumption that each data point is generated independently of each others. With that we can write

$$\Pr(Y_i = y_i, i = 1, \cdots, n) = \prod_{i=1}^{n} \Pr(Y_i = y_i)$$

Estimation in Simple Logistic Regression

With the assumption that the observations are independent, we get

$$\Pr(Y_i = y_i, i = 1, \cdots, n) = \prod_{i=1}^{n} \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1 - y_i}$$

$$= \prod_{i=1}^{n} \frac{(e^{\beta_0 + \beta_1 x_i})^{y_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Taking logarithms gives that

$$\ln \Pr(Y_i = y_i, i = 1, \cdots, n) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln \left(1 + e^{\beta_0 + \beta_1 x_i} \right)$$

Estimation in Simple Logistic Regression

From

$$\ln \mathrm{Pr}(Y_i = y_i, i = 1, \cdots, n) = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln\left(1 + e^{\beta_0 + \beta_1 x_i}\right)$$

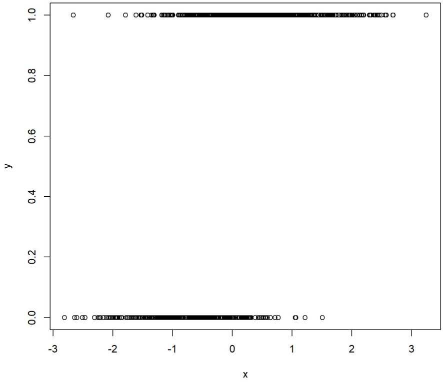
the maximum likelihood estimates can now be obtained by numerically finding the values of β_0 and β_1 that maximize the preceding likelihood.

This is not easy to do, and we typically use specialized software to obtain the estimates.

Logistic Regression in *R*

We now look at how to perform logistic regression in *R*: Let us import the simulated data from data set LR1.csv and plot the data.

```
LogisticRegression<-read.csv("LR1.CSV", header=TRUE, sep=",")
x<-LogisticRegression$x
y<-LogisticRegression$y
plot(x,y)
```



Logistic Regression in *R*

We use the following set of instructions to run logistic regression in *R*.

```
model1<-glm(y~x, data=LR, family="binomial")
summary(model1)
```

The *R*-output is given below:

```
Call:
glm(formula = y ~ x, family = "binomial", data = LR)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8653  -0.6338   0.3056   0.6863   2.9612

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.03341    0.09365   11.04   <2e-16 ***
x            2.03132    0.13508   15.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1287.31  on 999  degrees of freedom
Residual deviance:  857.34  on 998  degrees of freedom
AIC: 861.34

Number of Fisher Scoring iterations: 5
```

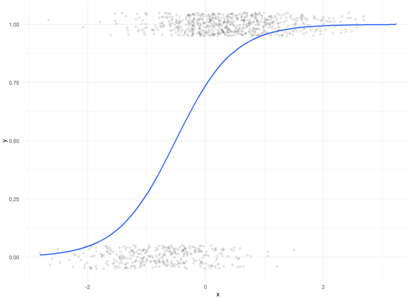
Logistic Regression in *R*

We can also use the following plotting instructions to obtain the characteristic logistic regression curve placed on top of that binary data.

```
library(tidyverse)
theme_set(theme_minimal())
LR<-read_csv("LR1.csv")

ggplot(LR, aes(x=x, y=y))+geom_jitter(height = 0.05, alpha=0.5)


ggplot(LR, aes(x=x, y=y))+geom_jitter(height = 0.05, alpha=0.1)+
  geom_smooth(method="glm", method.args=list(family="binomial"), se=FALSE)
```



Making Prediction

The $\text{predict}()$ function can be use to predict the probabilities of the form

$$\Pr(Y = 1|\mathbf{X}).$$

For that, we use $\text{type} =$ "response" as shown below:

```r
head(predict(model1, type="response"))
```

The *R*-output shows the predicted probabilities for the first 6 data.

```
        1         2         3         4         5         6 
0.4737492 0.6378032 0.9852208 0.7643417 0.7851664 0.9891990 
```

Making Prediction

Suppose you would like to obtain the prediction of the following probabilities

$$\Pr(Y = 1|X).$$

for $X = -0.000142, 0.45, 1.33, -0.56.$

The following input in *R*

```
newdata1 <- data.frame(x=c(-0.000142,0.45,1.33,-0.56))
predict(model1, newdata=newdata1, type="response")
```

will gives the requested prediction.

```
       1        2        3        4 
0.737520 0.875174 0.976685 0.473990 
```

Interpret the Model Parameters

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.03341    0.09365   11.04   <2e-16 ***
x            2.03132    0.13508   15.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

First notice from the *R*-output above that x does have a significant contribution to our response variable. Also, β_0 is not equal to zero, with any reasonable significant level.

Next, we can see that the fitted model is given by

$$\log(\text{odds}) = 1.03341 + 2.03132x$$

How do we make sense of this?

Interpret the Model Parameters

To answer the question, suppose we have

$$\ln\left(\frac{\pi}{1-\pi}\right)(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

suppose we let $x = x_0$ and consider the following set of equations

$$\ln\left(\frac{\pi}{1-\pi}\right)(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\ln\left(\frac{\pi}{1-\pi}\right)(x_0+1) = \hat{\beta}_0 + \hat{\beta}_1 (x_0+1)$$

Subtracting these two equation to obtain

$$\ln\left(\frac{\pi}{1-\pi}\right)(x_0+1) - \ln\left(\frac{\pi}{1-\pi}\right)(x_0) = \hat{\beta}_1$$

Interpret the Model Parameters

Now,

$$\ln\left(\frac{\pi}{1-\pi}\right)(x_0+1) - \ln\left(\frac{\pi}{1-\pi}\right)(x_0) = \hat{\beta}_1$$

let us write the equation above as

$$\ln\left(\mathrm{Odds}\right)(x_0+1) - \ln\left(\mathrm{Odds}\right)(x_0) = \hat{\beta}_1$$

hence we see that

$$\ln\left(\frac{\mathrm{Odds}(x_0+1)}{\mathrm{Odds}\ (x_0)}\right) = \hat{\beta}_1$$

or equivalently

$$\frac{\mathrm{Odds}(x_0+1)}{\mathrm{Odds}\ (x_0)} = e^{\hat{\beta}_1}$$

Interpret the Model Parameters

Therefore back to

$$\log(\text{odds}) = 1.03341 + 2.03132x$$

with $e^{2.03132} \approx 7.624$ and

$$\frac{\text{Odds}(x_0+1)}{\text{Odds}\ (x_0)} = e^{\hat{\beta}_1}$$

we can say the following:

For a one-unit increase from x_0 to $x_0 + 1$, we expect

$$\frac{\text{Odds}(x_0+1) - \text{Odds}(x_0)}{\text{Odds}(x_0)} \times 100\% = (e^{\hat{\beta}_1} - 1) \times 100\% = 662.4\%$$

a 662.4% increase in the odds (compare to $\text{Odds}(x_0)$).

Interpret the model parameters-multiple predictor variables

Suppose Y denote whether a student get admitted to attend Illinois Tech. While

- X_1 denote the math scores (0-100)
- X_2 denote the reading scores (0-100)

of a particular standardize exam. Suppose we collected n-data and

run a logistic regression to obtain

$$\ln\left(\frac{\pi}{1-\pi}\right) = 0.13 + 0.0456X_1 + 0.032X_2$$

where

$$\pi = \Pr(Y = 1|X_1 = x_1, X_2 = x_2).$$

How do we make sense of the regression coefficients?

Interpret the model parameters-multiple predictor variables

From

$$\ln\left(\frac{\pi}{1-\pi}\right) = 0.13 + 0.0456X_1 + 0.032X_2$$

We can say the following:

Holding the reading scores at a fixed value, the odds of being admitted to Illinois Tech is 4.67% higher for a one-unit increase in the math scores since

$$\frac{\mathrm{Odds}(x_0+1)-\mathrm{Odds}(x_0)}{\mathrm{Odds}(x_0)} \times 100\% = (e^{0.0456}-1)\times 100\% = 4.67\%$$

Interpret the model parameters-multiple predictor variables

Similarly, from

$$\ln\left(\frac{\pi}{1-\pi}\right) = 0.13 + 0.0456X_1 + 0.032X_2$$

We can say the following:

Holding the math scores at a fixed value, the odds of being admitted to Illinois Tech is 3.25% higher for a one-unit increase in the reading scores since

$$\frac{\mathrm{Odds}(x_0+1) - \mathrm{Odds}(x_0)}{\mathrm{Odds}(x_0)} \times 100\% = (e^{0.032} - 1) \times 100\% = 3.25\%$$