Week #7 - Multiple Logistic Regression Model

Cheng Peng

3/4/2021

Contents

1 Introduction

The general multivariable linear regression model is given below.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

where y is the response variable that is assumed to be a random variable and $\epsilon \to N(0, \sigma^2)$. This also implies that

$$E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

For a population with binary data, the underlying random variable can only takes exactly two values, say $Y = 1$ or $Y = 0$, and $P(Y = 1) = p$, then $E[Y] = 1 \times p + 0 \times (1 - p) = p$.

That is, the success probability is the expected value of the binary random variable. If we mimic the formulation of the linear regression model by setting

The simple linear regression model (also called log-odds regression model) is also formulated with the mean response $E[Y]$

$$\frac{E[Y]}{1 - E[Y]} = \beta_0 + \beta_1 x.$$

Let $g(t) = t/(1 - t)$ (also called logit function), the simple logistic regression is re-expressed as $g(E[Y]) = \beta_0 + \beta_1 x$.

2 Multiple Logistic Regression Model

Let Y be the binary response variable and $\{x_1, x_2, \cdots, x_n\}$ be the set of predictor variables. The multiple logistic regression model is defined as

$$\frac{E[Y]}{1 - E[Y]} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The success probability function

$$p(x_1, x_2, \cdots, x_k) = P(Y = 1 | x_1, x_2, \cdots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

2.1 Data Layout

Table 1: Data set layout for multiple logistic regression model

Y	X1	X2		Xk
Y1	X11	X21	...	Xk1
Y2	X12	X22	...	Xk2
...	...	...	...	...
Yn	X1n	X2n	...	Xkn

The regression coefficients are estimated using likelihood

2.2 Interpretation of Regression Coefficients

If we use numerical coding 1 = "success" and 0 = "failure", the simple logistic regression model can be explicitly expressed in the following form

$$\log \frac{P(Y = 1|x)}{1 - P(Y = 1|x)} = \beta_0 + \beta_1 x.$$

The expression $P(Y = 1|x)$ highlights that the success probability is dependent on the predictor variable x.

The regression coefficients of the simple logistic regression model have a meaningful interpretation.

- Intercept β_0 is the baseline log-odds of success. In other words, if the success probability is not impacted by any factors, β_0 is the log odds of success of the homogeneous population.
- the slope parameter β_1 is called log odds ratio of two categories corresponding to x and $x + 1$. To see this, denote $p_x = P(Y = 1|x)$ and $p_{x+1} = P(Y = 1|x + 1)$, then

$$\log \frac{p_x}{1 - p_x} = \beta_0 + \beta_1 x \quad \text{and} \quad \log \frac{p_{x+1}}{1 - p_{x+1}} = \beta_0 + \beta_1 (x + 1).$$

Taking the difference of the above two equations, we have

$$\log \frac{p_{x+1}}{1 - p_{x+1}} - \log \frac{p_x}{1 - p_x} = \beta_1$$

Therefore,

$$\log \frac{p_{x+1}/(1-p_{x+1})}{p_x/(1-p_x)} = \beta_1,$$

that is, β_1 is the ratio of log odds of success in two sub-populations.

2.3 Use of Simple Logistic Regression Model

We first express the success probability with respect to the predictor variable x in the following

$$P(Y=1|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}.$$

Using the above expression, we can of the following

- **Association Analysis** - if $\beta_1 \neq 0$, then the success probability is impacted by the predictor variable x. Note that, this is a non-linear association.
- **Predictive Analysis** - predicting the success probability for a given new value of the predictor variable. That is, $\widehat{P(Y=1|x_{new})} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{new}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{new}}}$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from the data.
- **Classification Analysis** - predicting the status of success. That is, for a given new value of predictor variable, we predict the value of Y through $P(Y=1|x_{new})$.
  - To predict whether the value of Y is "success" or "failure", we need to identify the cut-off probability to determine the value of Y.
  - The predicted model can be used in an **intervention analysis** - this means values of X can alter the value of Y. This is commonly used in the clinical study. For example, an effective treatment (X) can permanently cure a disease (Y).
  - The predicted model can be used for membership classification. For example, the response value is the gender (Y) of a car buyer at a car dealer, the predictor variable is the purchase status (X). If a customer bought a car from the dealer, the fitted model can identify whether the customer is a man or woman. This is apparently different from the intervention analysis since purchase status cannot change the gender (Y) of the customer.

2.4 Parameter Estimation

In linear regression models, both likelihood and least square methods can be used for estimating the coefficients of linear regression models. Both methods yield the same estimates. However, in the logistic regression model, we can only use the likelihood methods to estimate the regression coefficients.

Let $\{(y_1,x_1),(y_2,x_2),\cdots,(y_n,x_n)\}$ be a random sample taken from a binary population associated with Y. x is a nonrandom predictor variable associated with Y. The logistic model is defined to be

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

Since Y_i is a Bernoulli random variable with success probability p_x. We use numerical coding: 1 = "success" and 0 = "failure". The likelihood function of (β_0,β_1) is given by

$$L(\beta_0,\beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}\left[1-p(x_i)\right]^{1-y_i} = \prod_{i=1}^{n}\left[\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right]^{y_i} \times \left[\frac{1}{1+e^{\beta_0+\beta_1 x}}\right]^{1-y_i}$$

The maximum likelihood estimate (MLE) of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, maximizes the above likelihood. We will use the R build-in function **glm()** to find the MLE of the parameters and related statistics.

2.5 Model Assumptions and Diagnostics

In linear regression, we assume the response variable follows a normal distribution with a constant variance. With this assumption, several effective diagnostic methods were developed based on the residual analysis. In logistic regression, we don't have many diagnostic methods. However, several likelihood-based goodness of fit metrics such as AIC and deviance can be used for comparing the performance of candidate models.

More technical discussion of diagnostics with the left to the future specialized courses in generalized linear regression models.

2.6 Concluding Remarks

We only introduced the basic logistic regression modeling in this note. some important topics you may want to study but are not mentioned in this note are

- Logistic regression as a machine learning algorithm for predictive modeling.
- logistic regression model with a large number of predictor variables - regularized logistic regression.
- Performance metrics based on prediction errors.

3 Case Study

In this case study, we use the well-known iris data set. The original data set has one categorical variable - Species.

3.1 Data and Variable Descriptions

The original data has 4 numerical variables that reflect the sizes of iris flowers. There three different iris flowers in the data set, each iris flower has 50 observations. I will subset the data by including two iris species: "Iris-setosa" and "Iris-versicolor" and use Species as the binary response variable in this case study. We choose Sepal length (in cm) as the predictor variable to build a simple logistic regression model.

```
iris = read.csv("https://stat321.s3.amazonaws.com/w06-iris-csv.csv", header = TRUE)
binary.iris = iris[-which(iris$Species == "Iris-virginica"),]  # final data
```

3.2 Research Question

The objective of this case study is to explore the association between the species and sepal length.

3.3 Building the Simple Logistic Regression

Since the simple logistic regression contains only on continuous variable of a binary categorical variable as the predictor variable, no

```
s.logit = glm(as.factor(Species) ~ SepalLengthCm,  # as.factor() converts a categorical
                                                   # variable to a factor variable.
          family = binomial(link = "logit"),  # family is the binomial, logit(p) = log(p/(1-p))!
          data = binary.iris)      # the data frame is a subset if the original iris data
summary(s.logit)
```

```
##
## Call:
## glm(formula = as.factor(Species) ~ SepalLengthCm, family = binomial(link = "logit"),
##     data = binary.iris)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -2.05501  -0.47395  -0.02829   0.39788   2.32915
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -27.831      5.434  -5.122 3.02e-07 ***
## SepalLengthCm    5.140      1.007   5.107 3.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.629  on 99  degrees of freedom
## Residual deviance:  64.211  on 98  degrees of freedom
## AIC: 68.211
##
## Number of Fisher Scoring iterations: 6
```

Note that, as.factor() defines a two-level factor variable. The confidence interval of the log-odds ratio is given by the following code.

```
kable(confint(s.logit), caption = "The confidence interval of log odds ratios")
```

```
## Waiting for profiling to be done...
```

Table 2: The confidence interval of log odds ratios

	2.5 %	97.5 %
(Intercept)	-40.126184	-18.553199
SepalLengthCm	3.421613	7.415508

```
# Odds ratio
kable(exp(coef(s.logit)),caption = "Odds Ratios")
```

Table 3: Odds Ratios

	x
(Intercept)	0.0000
SepalLengthCm	170.7732

The odds ratio and the associated 95% confidence intervals are given below.

```
## odds ratios and 95% CI
kable(exp(cbind(OR = coef(s.logit), confint(s.logit))),caption = "Odds ratio and confidence intervals")
```

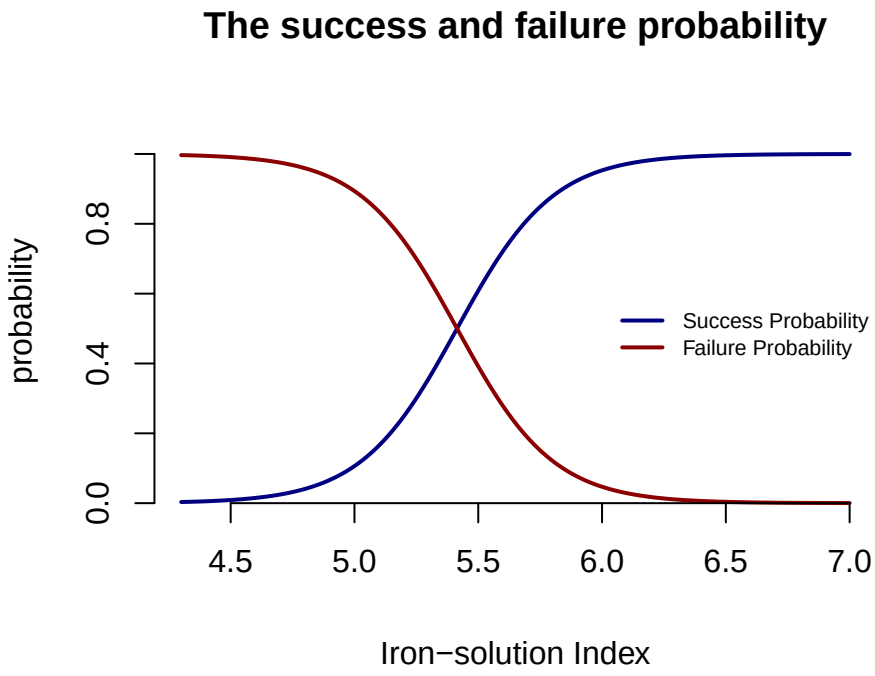
```
## Waiting for profiling to be done...
```

Table 4: Odds ratio and confidence intervals

	OR	2.5 %	97.5 %
(Intercept)	0.0000	0.00000	0.000
SepalLengthCm	170.7732	30.61878	1661.554

The success probability curve (so-called S curve) is given below.

```
sepal.length = range(binary.iris$SepalLengthCm)
x = seq(sepal.length[1], sepal.length[2], length = 200)
beta.x = coef(s.logit)[1] + coef(s.logit)[2]*x
success.prob = exp(beta.x)/(1+exp(beta.x))
failure.prob = 1/(1+exp(beta.x))
ylimit = max(success.prob, failure.prob)
##
plot(x, success.prob, type = "l", lwd = 2, col = "navy",
     main = "The success and failure probability",
     ylim=c(0, 1.1*ylimit),
     xlab = "Iron-solution Index",
     ylab = "probability",
     axes = FALSE)
lines(x, failure.prob,lwd = 2, col = "darkred")
axis(1, pos = 0)
axis(2)
legend(6.0, 0.6, c("Success Probability", "Failure Probability"), lwd = rep(2,2),
       col = c("navy", "darkred"), cex = 0.7, bty = "n")
```



3.4 Conclusion

The