Week #6 - Simple Logistic Regression Model

Due: 02/28/2021

2/27/2021

Table of Contents

[Introduction 1](#_Toc65677282)

[Simple Logistic Regression Model 2](#_Toc65677283)

[Model Structure 2](#_Toc65677284)

[Interpretation of Regression Coefficients 3](#_Toc65677285)

[Use of Simple Logistic Regression Model 4](#_Toc65677286)

[Parameter Estimation 5](#_Toc65677287)

[Model Assumptions and Diagnostics 5](#_Toc65677288)

[Concluding Remarks 5](#_Toc65677289)

[Case Study 6](#_Toc65677290)

[Data and Variable Descriptions 6](#_Toc65677291)

[Research Question 7](#_Toc65677292)

[Building the Simple Logistic Regression 7](#_Toc65677293)

[Conclusion 11](#_Toc65677294)

# Introduction

In the previous few modules, we reviewed the linear regression model systematically and also introduce nonparametric bootstrap methods to estimate the confidence intervals of regression coefficients.

The general formulation of a linear regression model is given below.

where is the response variable that is assumed to be a random variable. , , , are predictor variables that are assumed to be non-random. is the term of random error of the regression. It is assumed to have a normal distribution with mean 0 and constant variance. That is, . Since . This also implies that

In other words, the response variable in linear regression models MUST be continuous and, theoretically speaking, a normal random variable with its mean depending on predictor variables.

From the distribution of , we can see that

Therefore, a linear regression model assesses how predictor variables impact the mean of the response variable. The regression coefficients reflect the association between the response and the corresponding predictor variables.

In this module, we study a new regression model with a binary response variable that takes on exactly one of two possible values such as “success” v.s. “failure”, “diseased” v.s. “disease-free”, etc.

For a binary population, we are interested in the proportion of one of the two values. For example, if we study a disease of a certain population, the primary interest in the prevalence of a disease - the proportion of subjects in the population who had the disease. We also know that the relative frequency (proportion) of the disease can be used to estimate the disease probability of the population. A natural and practical question is that whether the disease probability is impacted by some factors - this is a repression problem. In the next section, we formulate this type of regression focusing on the binary response.

# Simple Logistic Regression Model

In this section, we focus on a special case by considering a single numerical or a binary categorical predictor in the model.

## Model Structure

Let be a binary variable that takes on exactly one of the two possible values, say “success” or “failure”. Assume the proportion of “success” to be . That is, P(Y = “success”) = p. Let be a factor that may impact the success probability.

Recall that linear regression is formulated by . Next, we calculate the expected value of the binary response variable.

In the rest of this section, we assume = binary variable that takes on either “success” or “failure”. = a factor that impacts the success probability of success. Further P(Y=“success”) = p. Clearly, is a Bernoulli random variable. Since we are interested in the success probability, we perform a numerical coding by setting “success” to 1 and “failure” to 0. By definition

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 above formulation is inappropriate since the right-hand side of the equation can take on any real value while the left-hand side equation can take on only values in the interval [0,1].

One of the most important criteria of being a good statistical model is its interpretability. We cannot link the success probability with the linear function of the predictor. Next, we link the **odds of success** () with the linear function of the predictor variable in the following.

The above expression shows the association between the odds of success, , and the predictor variable . It is interpretable. But the odds of success takes on the value in . However, , it is reasonable to explore the association between the logarithm of odds of success and the predictor variable . That is,

The above model is called the **simple logistic regression model**.

**Remarks on the model formulation**

* The above logistic regression model was formulated based on the interpretability of the transformed mean response, , and its linear association with the predictor variable.
* Mathematically, there are different transformations of the mean response to define a family of regression models . Clearly, the above simple logistic regression model is a special member of this family - generalized linear regression models. There are two other less famous but commonly used in some specific fields: complementary log-log model and probit model.

## 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

The expression highlights that the success probability is dependent on the predictor variable .

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

* Intercept is the baseline log-odds of success. In other words, if the success probability is not impacted by any factors, is the log odds of success of the homogeneous population.
* the slope parameter is called log odds ratio of two categories corresponding to and . To see this, denote and , then

$$ \log \frac{p\_x}{1-p\_x} = \beta\_0 + \beta\_1x\hspace{3mm} \mbox{and} \hspace{3mm} \log \frac{p\_{x+1}}{1-p\_{x+1}} = \beta\_0 + \beta\_1(x+1). $$

Taking the difference of the above two equations, we have

Therefore,

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

## Use of Simple Logistic Regression Model

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

Using the above expression, we can of the following

* **Association Analysis** - if , then the success probability is impacted by the predictor variable . 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, , where and 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 through .
  + To predict whether the value of is “success” or “failure”, we need to identify the cut-off probability to determine the value of .
  + The predicted model can be used in an **intervention analysis** - this means values of can alter the value of . This is commonly used in the clinical study. For example, an effective treatment () can permanently cure a disease ().
  + The predicted model can be used for membership classification. For example, the response value is the gender () of a car buyer at a car dealer, the predictor variable is the purchase status (). 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 () of the customer.

## 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 be a random sample taken from a binary population associated with . is a nonrandom predictor variable associated with . The logistic model is defined to be

Since is a Bernoulli random variable with success probability . We use numerical coding: 1 = “success” and 0 = “failure”. The likelihood function of is given by

The maximum likelihood estimate (MLE) of and , denoted by and , maximizes the above likelihood. We will use the R build-in function **glm()** to find the MLE of the parameters and related statistics.

## 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.

## 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.
* a logistic regression model with a large number of predictor variables - regularized logistic regression.
* Performance metrics based on prediction errors.

# Case Study

The diabetes data set in this case-study contains 768 observations on 9 variables. The data set is available in the UCI machine learning data repository. R library {mlbench} has two versions of this data. The data set contains a significant number of missing values.

## Data and Variable Descriptions

There are 9 variables in the data set.

1. **pregnant**: Number of times pregnant
2. **glucose**: Plasma glucose concentration (glucose tolerance test)
3. **pressure**: Diastolic blood pressure (mm Hg)
4. **triceps**: Triceps skin fold thickness (mm)
5. **insulin**: 2-Hour serum insulin (mu U/ml)
6. **mass**: Body mass index (weight in kg/(height in m)^2)
7. **pedigree**: Diabetes pedigree function
8. **age**: Age (years)
9. **diabetes**: Class variable (test for diabetes)

I load the data from R **library{mlbench}** in the following code.

library(mlbench)  
data(PimaIndiansDiabetes2) # load the data to R work-space  
diabetes.0 = PimaIndiansDiabetes2 # make a copy of the data for data cleansing   
diabetes = na.omit(diabetes.0) # Delete all records with missing components  
y0=diabetes$diabetes  
diabete.01 = rep(0, length(y0)) # define a 0-1 to test which probability is used in glm()  
diabete.01[which(y0=="pos")] = 1  
diabetes$diabetes.01 = diabete.01  
# head(diabetes)

For convenience, I delete all records with missing values and keep only the records with complete records in this case study. The final analytic data set has 392 records.

## Research Question

Many studies indicated that the body mass index (BMI) is a more powerful risk factor for diabetes than genetics. The objective of this case study is to explore the association between BMI and diabetes.

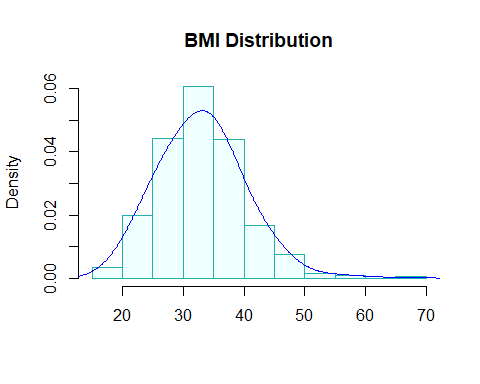
The general interpretation of BMI for adults is given below:

* **underweight**: < 18.5
* **Normal and Healthy Weight**: [18.5, 24.9]
* **Overweight**: [25.0, 29.9]
* **Obese**: > 30.0

## Building the Simple Logistic Regression

Since we only study a simple logistic regression model, only one predictor variable is included in the model. We first perform exploratory data analysis on the predictor variable to make sure the variable is not extremely skewed.

ylimit = max(density(diabetes$mass)$y)  
hist(diabetes$mass, probability = TRUE, main = "BMI Distribution", xlab="",   
 col = "azure1", border="lightseagreen")  
 lines(density(diabetes$mass, adjust=2), col="blue")



Since the simple logistic regression contains only one continuous variable of a binary categorical variable as the predictor variable, no there is no issue of potential imbalance. We will not transform BMI and fit a logistic regression directly to the data.

s.logit = glm(diabetes ~ mass,   
 family = binomial(link = "logit"), # family is the binomial, logit(p) = log(p/(1-p))!  
 data = diabetes) # the data frame is a subset of the original iris data  
# summary(s.logit)

Note that the response variable is a binary factor variable, Ruses alphabetical order to define the level of the factor variable. In our case, “neg” = 0 and “pos” = 1. The “success” probability is defined to be P(diabetes = “pos”). The simple logistic regression is fitted in the following.

The summary of major statistics is given below.

model.coef.stats = summary(s.logit)$coef # output stats of coefficients  
conf.ci = confint(s.logit) # confidence intervals of betas

## Waiting for profiling to be done...

sum.stats = cbind(model.coef.stats, conf.ci.95=conf.ci) # rounding off decimals  
kable(sum.stats,caption = "The summary stats of regression coefficients")

The summary stats of regression coefficients

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) | 2.5 % | 97.5 % |
| (Intercept) | -3.6061432 | 0.5917329 | -6.094208 | 0e+00 | -4.8064211 | -2.4817472 |
| mass | 0.0863295 | 0.0170535 | 5.062276 | 4e-07 | 0.0538266 | 0.1208267 |

From the above table, we can see that BMI is positively associated with the status of diabetes since with a p-value close to 0. The 95% confidence interval [0.0538, 0.1208]. This also supports the results of the research in the literature.

It is more common to interpret the association results from a practical perspective using the odds ratio. Next, we convert the estimated regression coefficients in the odds ratio.

# Odds ratio  
model.coef.stats = summary(s.logit)$coef  
odds.ratio = exp(coef(s.logit))  
out.stats = cbind(model.coef.stats, odds.ratio = odds.ratio)   
kable(out.stats,caption = "Summary Stats with Odds Ratios")

Summary Stats with Odds Ratios

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) | odds.ratio |
| (Intercept) | -3.6061432 | 0.5917329 | -6.094208 | 0e+00 | 0.0271564 |
| mass | 0.0863295 | 0.0170535 | 5.062276 | 4e-07 | 1.0901655 |

The odds ratio associated with BMI is 1.09 meaning that as the BMI increases one unit, the odds of being tested positive in diabetes increases by about . This is a practically significant risk factor for diabetes.

Some global goodness-fo-fit measures are summarized in the following table.

## Other global goodness-of-fit  
dev.resid = s.logit$deviance  
dev.0.resid = s.logit$null.deviance  
aic = s.logit$aic  
goodness = cbind(Deviance.residual =dev.resid, Null.Deviance.Residual = dev.0.resid,  
 AIC = aic)  
kable(goodness, caption="Global Goodness-of-fit Measures")

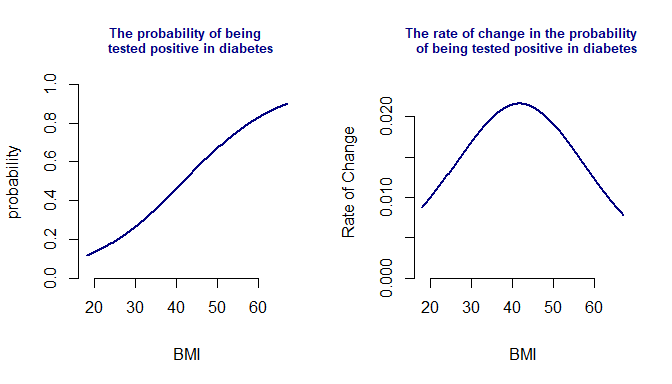
Global Goodness-of-fit Measures

|  |  |  |
| --- | --- | --- |
| Deviance.residual | Null.Deviance.Residual | AIC |
| 469.031 | 498.0978 | 473.031 |

Since the above global goodness-of-fit are based on the likelihood function, we don’t have other candidate models with corresponding likelihood at the same scale to compare in this simple logistic regression model, we will not interpret these goodness-of-fit measures.

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

bmi.range = range(diabetes$mass)  
x = seq(bmi.range[1], bmi.range[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)  
##  
beta1 = coef(s.logit)[2]  
success.prob.rate = beta1\*exp(beta.x)/(1+exp(beta.x))^2  
##  
##  
par(mfrow = c(1,2))  
plot(x, success.prob, type = "l", lwd = 2, col = "navy",  
 main = "The probability of being \n tested positive in diabetes",   
 ylim=c(0, 1.1\*ylimit),  
 xlab = "BMI",  
 ylab = "probability",  
 axes = FALSE,  
 col.main = "navy",  
 cex.main = 0.8)  
# lines(x, failure.prob,lwd = 2, col = "darkred")  
axis(1, pos = 0)  
axis(2)  
# legend(30, 1, c("Success Probability", "Failure Probability"), lwd = rep(2,2),   
# col = c("navy", "darkred"), cex = 0.7, bty = "n")  
##  
y.rate = max(success.prob.rate)  
plot(x, success.prob.rate, type = "l", lwd = 2, col = "navy",  
 main = "The rate of change in the probability \n of being tested positive in diabetes",   
 xlab = "BMI",  
 ylab = "Rate of Change",  
 ylim=c(0,1.1\*y.rate),  
 axes = FALSE,  
 col.main = "navy",  
 cex.main = 0.8  
 )  
axis(1, pos = 0)  
axis(2)



The left-hand side plot in the above figure is the standard **S curve** representing how the probability of positive test increases as the BMI increases. After diving deeper to see the rate of change in the probability of positive test, we obtain the curve on the right-hand side that indicates that the rate of change in the probability of positive test increases when BMI is less 40 and decreases when BMI > 40. The turning point is about 40.

# Conclusion

This note focuses on the structure and association analysis of the simple logistic regression model. The case study uses a real-world diabetes data set to illustrate the steps for carrying the simple logistic regression model.

In the following modules, we will discuss multiple logistic regression models that will include more modeling techniques.

Since logistic regression has been used as a standard machine learning algorithm for classification, I will use a standalone module to discuss this topic.