Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 11

Logistic regression

Logistic regression

- In linear regression modeling, the outcome variable is a continuous variable.
- When the outcome variable is categorical in nature, logistic regression can be used to predict the likelihood of an outcome based on the input variables.
- Although logistic regression can be applied to an outcome variable that represents multiple values, in this course we will focus on the case in which the outcome variable is binary (e.g. true/false, pass/fail, or yes/no).

Logistic regression

- For example, a logistic regression model can be built to determine if a person will or will not purchase a new automobile in the next 12 months.
- The training set could include input variables for a person's age, income, and gender as well as the age of an existing automobile.
- The training set would also include the outcome variable on whether the person purchased a new automobile over a 12-month period.
- The logistic regression model provides the likelihood or probability of a person making a purchase in the next 12 months.

Logistic function

- Logistic regression is based on the logistic function

$$f(z) = \frac{\exp(z)}{1+\exp(z)}, \text{for } -\infty < z < \infty.$$

- Note that as $z \to \infty$, $f(z) \to 1$.
- Also, as $z \to -\infty$, $f(z) \to 0$.

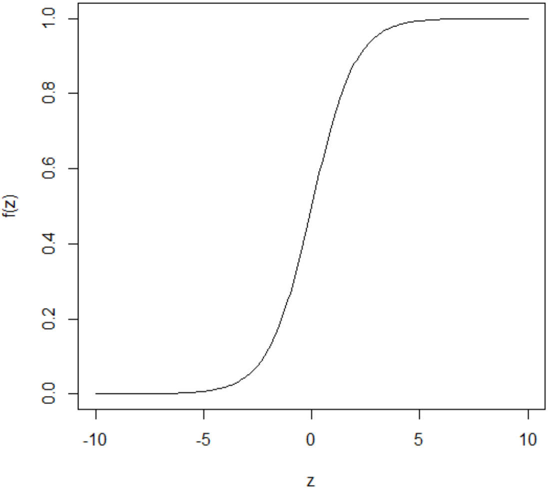
Logistic function

- We can plot the logistic function in R to visualize these properties.
-

```
logistic = function(z) {
  exp(z)/(1+exp(z))
}

z = seq(-10,10,0.1);
plot(z, logistic(z), xlab="z", ylab="f(z)", lty=1,
     type='l')
```

Logistic function



Logistic Regression

- In any proposed model, to predict the likelihood of an outcome, z needs to be a function of the input or feature variables X.
- In logistic regression, z is expressed as a linear function of the input variables:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

- Then in logisitc regression, $P(Y = 1|X_1, X_2, ..., X_p)$ can be expressed as

$$P(Y = 1|X_1, X_2, ..., X_p) \\ = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}$$

Logistic Regression

- As a simple example, suppose there is only a single feature variable X. Then the model for logistic regression is:

$$\pi(X) = P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

- Assume the true parameter values are $\beta_0 = -5$, $\beta_1 = 1$.
- We can plot the function $\pi(X)$ versus different values for the variable Xgreatest curvature at -5 and 1

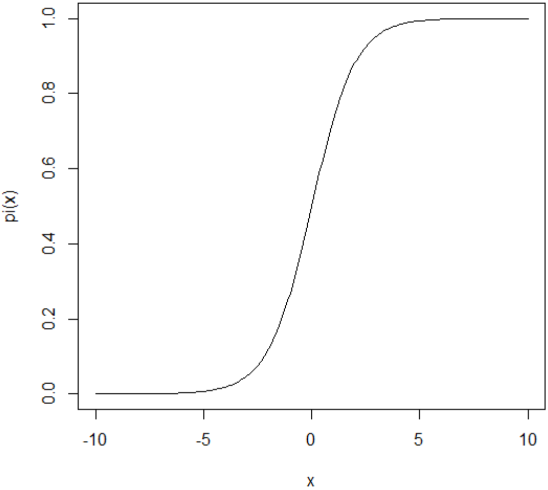
Logistic function

- Code in R:
-

```
logistic = function(z) {
  exp(z)/(1+exp(z))
}

beta0 = -5
beta1 = 1
x = seq(-5,15,0.1);
plot(z, logistic(beta0+beta1*x), xlab="x", ylab="pi(x)", lty=1, type='l')
```

Logistic function



Logistic function

- Just like simple linear regression, in logistic regression the parameters β_0 andβ_1 need to be estimated based on training data.
- Instead of the method of *least squares*, parameter estimation in logistic regression is based on the method called *maximum likelihood estimation* (MLE).
- We will start with a gentle introduction to *MLE*; more extensive coverage of this method is available in courses such as ST2131/MA2216 Probability and ST2132 Mathematical Statistics.

Introduction to MLE

- We will first illustrate MLE with a coin toss example.
- Suppose we toss a (possibly unfair) coin three times; the outcome Y for each toss is head or tail.
- Let the probability of the coin coming up a head (Y=1) be p.
- Then the probability of the coin coming up a tail (Y=0) is $1 - p$.
- In summary, $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$.
- p is unknown, and we wish to estimate its value based on data from the three tosses. training data

Introduction to MLE

- Suppose the tosses are "independent" and the same coin is used each time.
- We observe the first toss to come up a head and the subsequent two tosses come up as tails.
- Then the *likelihood function* of p given the observed data is

$$L(p) = p \times (1 - p) \times (1 - p) = p(1 - p)^2$$

Introduction to MLE

- Suppose the tosses are "independent" and the same coin is used each time.
- We observe the first toss to come up a head and the subsequent two tosses come up as tails.
- Then the *likelihood function* of p given the observed data is

$$L(p) = p \times (1 - p) \times (1 - p) = p(1 - p)^2$$

- To estimate a parameter, the method of *maximum likelihood* chooses the parameter value that makes L as large as possible.

Introduction to MLE

- The *likelihood function* of p given the observed data is

$$L(p) = p \times (1-p) \times (1-p) = p(1-p)^2$$

- To estimate a parameter, the method of *maximum likelihood* chooses the parameter value that makes L as large as possible.
-

$$\begin{aligned}\frac{\partial L(p)}{\partial p} &= \frac{\partial}{\partial p}\left(p - 2p^2 + p^3\right) = 1 - 4p + 3p^2 \\ &= 3\left(p^2 - \frac{4}{3}p + \frac{1}{3}\right) = 3\left(p - \frac{1}{3}\right)(p-1) = 0.\end{aligned}$$

Introduction to MLE

- The *likelihood function* of p given the observed data is

$$L(p) = p \times (1-p) \times (1-p) = p(1-p)^2$$

- To estimate a parameter, the method of *maximum likelihood* chooses the parameter value that makes L as large as possible.
-

$$\frac{\partial^2 L(p)}{\partial p^2} = 6p - 4$$

- For $p = \frac{1}{3}$, $6 \times \frac{1}{3} - 4 < 0$; For $p = 1$, $6 - 4 > 0$
- $\rightarrow$ Maximum for $P(p)$ is achieved at $\hat{p}_{mle} = \frac{1}{3}$

Introduction to MLE

- $\hat{p}_{mle} = \frac{1}{3}$ is the *maximum likelihood* estimate for p based on the results of the three tosses.
- We observe that it may also agree with our intuitive estimate for the probability that a toss comes up as head.

Introduction to MLE

- As another example, suppose we toss the coin three times and we observe the first two tosses to come up as heads and the last toss come up as tail.
- Suppose the tosses are "independent" and the same coin is used each time.
- Then the *likelihood function* of p given the observed data is

$$L(p) = p \times p \times (1 - p) = p^2(1 - p)$$

Introduction to MLE

- To estimate a parameter, the method of *maximum likelihood* chooses the parameter value that makes L as large as possible.
-

$$\frac{\partial L(p)}{\partial p} = \frac{\partial}{\partial p}\left(p^2 - p^3\right) = 2p - 3p^2$$
$$= p(2 - 3p) = 0.$$

Introduction to MLE

- $$\frac{\partial^2 L(p)}{\partial p^2} = 2 - 6p$$
- For $p = \frac{2}{3}$, $2 - 6 \times \frac{2}{3} < 0$; For $p = 0$, $2 > 0$
- $\rightarrow$ Maximum for $P(p)$ is achieved at $\hat{p}_{mle} = \frac{2}{3}$

Introduction to MLE

- In general, suppose we toss the coin n times. For the i^{th} toss, let $y_i = 1$ if it comes up a head and $y_i = 0$ if it comes up a tail.
- Then the likelihood function given the observed data is:

$$L(p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}.$$

y(i)=1, p left

y(i)=0, 1-p left

Introduction to MLE

- To obtain the *maximum likelihood* estimate, it is convenient to work with the logarithm of L rather than with L itself.
- Then the log- likelihood function given the observed data is:

$$\ln L(p) = \sum_{i=1}^{n} y_i \ln(p) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - p).$$

- Note that $n_1 = \sum_{i=1}^{n} y_i$ is just the total number of heads, and $n_0 = \sum_{i=1}^{n} (1 - y_i)$ is the total number of tails, where $n_1 + n_0 = n$.
- So the log-likelihood function is equivalently:

$$\ln L(p) = n_1 \ln(p) + n_0 \ln(1 - p).$$

Introduction to MLE

- Differentiate the log-likelihood with respect to p yields:

$$\frac{n_1}{p} - \frac{n_0}{1-p} = 0$$

$$\rightarrow n_1(1-p) = n_0 p$$

$$\rightarrow \hat{p}_{mle} = \frac{n_1}{n_1 + n_0} = \frac{n_1}{n}.$$

- Technically, we may still need to verify that we are at a maximum (rather than a minimum) by seeing if the second derivative is negative.

Introduction to MLE

- Now again suppose we toss the coin n times.
- For the i^{th} toss, a dollar coin $(x_i = 1)$ or a non-dollar coin $(x_i = 0)$ is used.
- For the i^{th} toss, let $y_i = 1$ if it comes up a head and $y_i = 0$ if it comes up a tail.
- So we observe data $x = c(x_1, x_2, ..., x_n)$ and $y = c(y_1, y_2, ..., y_n)$.
- How to incorporate the binary feature variable X into the MLE framework that we discussed?

Logistic Regression

- To incorporate the single binary feature variable X, let

$$\pi(X) = P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

so that

$$P(Y = 0|X) = 1 - \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} = \frac{1}{1 + \exp(\beta_0 + \beta_1 X)}$$

- Before estimating β_0 and β_1 via MLE, we need to look at the form of the likelihood function first:

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

Logistic Regression

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

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

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

Logistic Regression

- The log-likelihood function is:

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_i) - \ln(1 + \exp(\beta_0 + \beta_1 x_i))\}$$

- We will continue with estimation of β_0 and β_1 on Friday.
- We will look at simple examples of how this estimation procedure is carried out.

Review: Derivatives of the Exponential Function

- Recall that

$$\frac{d}{dx}e^x = e^x$$

  and

$$\frac{d}{dx}e^{f(x)} = e^{f(x)}\frac{d}{dx}f(x)$$

- For example,

$$\begin{aligned}\frac{d}{dx}e^{x^2} &= e^{x^2}\frac{d}{dx}x^2 \\ &= \left(e^{x^2}\right)2x\end{aligned}$$