APPLIED STATISTICS

Logistic Regression for Two-Category Response Variables and Its Estimation

Dr Tao Zou

Research School of Finance, Actuarial Studies & Statistics
The Australian National University

Last Updated: Tue Sep 26 13:52:35 2017

Overview

- Two-Category Response Variables
- Motivating Example
- Bianry Logistic Regression Model
- Estimation of Bianry Logistic Regression
- Prediction of a New Observation

References

1. **F.L. Ramsey and D.W. Schafer** (2012)
   Chapter 20 of *The Statistical Sleuth*

2. ANU STAT3015 Lecture Notes

3. The slides are made by **R Markdown**.
   http://rmarkdown.rstudio.com

Two-Category Response Variables

In numerous regression applications, the response variable of interest is a categorical variable taking two values.

In such situations the response can be represented by a binary indicator variable taking on values 0 and 1. For example:

- In a study on the effectiveness of a new drug, the response might be whether a given patient survived a 5-year period.
- In a study of home ownership, the response variable is whether a given individual owns a home.

Example: Anaesthetic Data

(Taken from STAT3015 notes.)

The potency of an anaesthetic agent is measured in terms of the minimum concentration at which at least 50% of patients exhibit no response to stimulation.

Thirty patients were given a particular anaesthetic at various predetermined concentrations for 15 minutes before a stimulus was applied.

The response variable was simply an indication as to whether the patient responded to the stimulus in any way.

"Response" is 1 if the patient responded to the stimulus.

R Code

```
setwd('~/Desktop/Research/AppliedStat2017/L9')
a=read.csv('anaesthetic.csv');a
```

```
##    Concentration Response
## 1            0.8        1
## 2            0.8        1
## 3            0.8        1
## 4            0.8        1
## 5            0.8        1
## 6            0.8        1
## 7            0.8        0
## 8            1.0        1
## 9            1.0        1
## 10           1.0        1
## 11           1.0        1
## 12           1.0        0
## 13           1.2        1
## 14           1.2        1
## 15           1.2        0
## 16           1.2        0
## 17           1.2        0
## 18           1.2        0
## 19           1.4        1
## 20           1.4        1
## 21           1.4        0
## 22           1.4        0
## 23           1.4        0
## 24           1.4        0
## 25           1.6        0
## 26           1.6        0
## 27           1.6        0
## 28           1.6        0
## 29           2.5        0
## 30           2.5        0
```

X Y

the number of Response = 1

$$\frac{1+1+1+1+1+1}{7} = \frac{6}{7}$$

sample mean of responses at $X = 0.8$

var

the proportion of Response = 1

[0, 1]

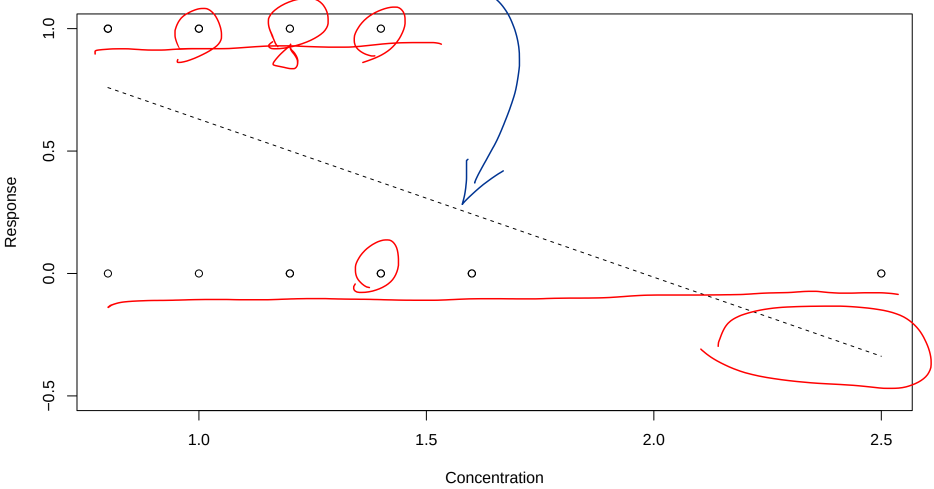
should be in [0, 1].

/L22

R Code (Con'd)

```
attach(a)
plot(Concentration, Response,ylim=c(-0.5,1))
fit=lm(Response~Concentration)
lines(Concentration,fit$fitted,lty=2)
```

dashed line



On this scale, a linear regresion does not seem appropriate.

Violation of Linear Regression Assumptions

Y: Response; X: Concentration.

1. Y not conform normality assumption, since Y only takes values of 0 and 1.

2. target

```
tapply(Response, Concentration,mean)
```

→ returns the means of target by Concentration

```
##       0.8         1       1.2       1.4       1.6       2.5
## 0.8571429 0.8000000 0.3333333 0.3333333 0.0000000 0.0000000
```

→ sample mean by different

Given $X = 0.8$, the sample mean of Y is 0.857; $= \frac{6}{7}$
given $X = 1.0$, the sample mean of Y is 0.800;
given $X = 1.2$, the sample mean of Y is 0.333;
given $X = 1.4$, the sample mean of Y is 0.333;
given $X = 1.6$, the sample mean of Y is 0.000;
given $X = 2.5$, the sample mean of Y is 0.000.

Based on data, the sample mean is actually the proportion that $Y = 1$ given $X = x$, and hence should be in the interval $[0, 1]$.

This indicates that the mean of Y given $X = x$ ($\mu\{Y|X = x\}$) should be in $[0, 1]$.
But in linear regression, $\mu\{Y|X = x\} = \beta_0 + \beta_1 x$ can take values outside of $[0, 1]$.

Violation of Linear Regression Assumptions (Con'd)

3.

```
tapply(Response, Concentration,var)
```

```
##       0.8         1       1.2       1.4       1.6       2.5
## 0.1428571 0.2000000 0.2666667 0.2666667 0.0000000 0.0000000
```

Given $X = 0.8$, the sample variance of Y is 0.143;
given $X = 1.0$, the sample variance of Y is 0.200;
given $X = 1.2$, the sample variance of Y is 0.267;
given $X = 1.4$, the sample variance of Y is 0.267;
given $X = 1.6$, the sample variance of Y is 0.000;
given $X = 2.5$, the sample variance of Y is 0.000.

The constant variance assumption is violated, $\sigma\{Y|X = x\}$ are not constant.

Problem 3 could be fixed using weighted regression. Problem 1 may not be a problem since LS estimates are robust to some non-normal distributions. Problem 2 is more problematic.

Generalised Linear Model (GLM)

The above example indicates that the mean of Y given $X = x$ (i.e., $\mu\{Y|X = x\}$) should be in the interval $[0, 1]$ for a binary response Y.

But in the linear regression, $\mu\{Y|X = x\} = \beta_0 + \beta_1 x$ can take values outside of $[0, 1]$.

So how about we find some transformation $h(\cdot)$ such that

$$\boxed{\mu\{Y|X = x\} = h(\beta_0 + \beta_1 x)} \in [0, 1] \text{ for sure?}$$

e.g. $h(v) = \dfrac{e^v}{1+e^v}$

$h^{-1}(u) = \log \dfrac{u}{1-u} =: g(u)$

Usually we consider the function $h(\cdot)$ to force that

$$\boxed{u = h(v) \Rightarrow v = h^{-1}(u) = g(u),} \text{ say,}$$

definition of inverse function

namely g is the inverse function of h. Also h is the inverse function of g, i.e., $h(v) = g^{-1}(v)$.

link fuction

Then

$$\beta_0 + \beta_1 x = h^{-1}(\mu\{Y|X = x\}) = g(\mu\{Y|X = x\}).$$

Generalised Linear Model (Con'd)

A generalised linear model (GLM) is a model where the mean of the response is related to the explanatory variables via the following relationship:

$$g\left(\mu\{Y|X_1, \cdots, X_k\}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

This relationship is linear in the parameters. The function $g(\cdot)$ is called the link function.

The choice of link function $g(\cdot)$ depends on the type of the response variable, and is not limited to a binary response Y.

In this lecture we introduce the link function for two-category resonse Y (in such situations the response can be represented by a binary indicator variable taking on values 0 and 1).

We call this proposed model with a specific link for two-category response: **binary logistic regression** model.

$\to g(\cdot)$

Other link functions lead to other GLMs, where in these cases the response is not necessarily binary.

Overview of This Course

	Continuous X + Categorical X
Continuous Y	MLR + Indicator Variables
Two-Category Y	**Binary Logistic Regression + Indicator Variables**

Binary Logistic Regression Model Assumptions

1. **Bernoulli distribution**: There is a Bernoulli distributed (sub)population of responses for given values of the explanatory variables $(X_1 = x_1, \cdots, X_k = x_k)$. That means if we let $X = (X_1, \cdots, X_k)$, the probability that $Y = 1$ given X is

$$\mathrm{P}(Y = 1|X) = \pi(X) \in [0, 1], \text{ and}$$

$$\mathrm{P}(Y = 0|X) = 1 - \mathrm{P}(Y = 1|X) = 1 - \pi(X).$$

$$\mu\{Y|X\} = 1 \times \mathrm{P}(Y = 1|X) + 0 \times \mathrm{P}(Y = 0|X) = \pi(X) \in [0, 1].$$

2. **Generalised Linearity**: The transformation of the mean of response falls on a linear function of the explanatory variables

$\Rightarrow \mu\{Y|X\} = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$

$$g(\mu\{Y|X\}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \text{ for } X = (X_1, \cdots, X_k),$$

where $g(u) = \log\{u/(1 - u)\}$, which is called logit link function.

Binary Logistic Regression Model Assumptions (Con'd)

Remark: the inverse function of the logit link function is

$$g^{-1}(v) = \frac{e^v}{1+e^v} \in [0,1]. \rightarrow \text{logistic function}$$

$\pi(X) = P(Y=1|X) =$

Then

$$\mu\{Y|X\} = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) \in [0,1],$$

which is consistent with the range $\mu\{Y|X\} = \pi(X) \in [0,1]$.

3. **Independence**: Observations

$$(X_{1,1}, \cdots X_{k,1}, Y_1),$$

$$\vdots$$

$$(X_{1,n}, \cdots X_{k,n}, Y_n),$$

are independent, where n is the sample size.

Binary Logistic Regression and Interpretation

Based on the above assumptions,

$$\begin{aligned} \mathrm{P}(Y=1|X) &= \mu\{Y|X\} = g^{-1}(\beta_0+\beta_1X_1+\cdots+\beta_kX_k) \\ &= \frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}} \end{aligned}$$

Then we compute

$$\frac{\mathrm{P}(Y=1|X)}{1-\mathrm{P}(Y=1|X)} = e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}$$

which is called odds that $Y=1$ given X.

$\boxed{\beta_0+\beta_1x_1+\cdots+\beta_kX_k} = 0$

- odds $=1$ means there is a 50% chance that $Y=1$ will occur $[\mathrm{P}(Y=1|X)=0.5]$.
- odds >1 means there is a better than 50% chance that $Y=1$ will occur $[\mathrm{P}(Y=1|X)>0.5]$.
- odds <1 means there is less than 50% ~~chance~~ that $Y=1$ will occur $[\mathrm{P}(Y=1|X)<0.5]$.

Hence, odds is another way to describe probability.

Binary Logistic Regression and Interpretation (Con'd)

hold other constant

Then

at

$$\text{odds} = \frac{\mathrm{P}(Y=1|X_1=x_1+1, X_2=x_2, \cdots, X_k=x_k)}{1-\mathrm{P}(Y=1|X_1=x_1+1, X_2=x_2, \cdots, X_k=x_k)} = e^{\beta_0+\beta_1(x_1+1)+\cdots+\beta_k x_k} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k} e^{\beta_1}, \text{ and}$$

at

$$\text{odds} = \frac{\mathrm{P}(Y=1|X_1=x_1, X_2=x_2, \cdots, X_k=x_k)}{1-\mathrm{P}(Y=1|X_1=x_1, X_2=x_2, \cdots, X_k=x_k)} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}.$$

With the other variables held constant, if X_1 is increased by 1 unit, the odds that $Y=1$ will change by a multiplicative factor of e^{β_1}.

Δ Δ β_1

Estimation of Binary Logistic Regression Parameters

For all generalised linear models, the method of least squares is replaced by the method of maximum likelihood estimation (MLE).

Consider the response $Y = y$,

$y = 1 \Rightarrow$

$\mathrm{P}(Y = 1|X) = \pi(X) = \{\pi(X)\}^1\{1-\pi(X)\}^{1-1} = \{\pi(X)\}^y\{1-\pi(X)\}^{1-y}$,

$y = 0 \Rightarrow$

$\mathrm{P}(Y = 0|X) = 1-\pi(X) = \{\pi(X)\}^0\{1-\pi(X)\}^{1-0} = \{\pi(X)\}^y\{1-\pi(X)\}^{1-y}$.

Hence,

$$\boxed{\mathrm{P}(Y = y|X) = \{\pi(X)\}^y\{1-\pi(X)\}^{1-y}.}$$

y can be both 0, 1.

It is worth noting that

logistic function

$$\boxed{\pi(X)} = \mu\{Y|X\} = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) =: p(\beta_0, \cdots, \beta_k), \text{ say.}$$

Since $\boxed{X = (X_1, \cdots, X_k)}$, we have

$$\begin{aligned}\mathrm{P}(Y = y|X_1, \cdots, X_k) &= \{\pi(X_1, \cdots, X_k)\}^y\{1-\pi(X_1, \cdots, X_k)\}^{1-y} \\ &= \{p(\beta_0, \cdots, \beta_k)\}^y\{1-p(\beta_0, \cdots, \beta_k)\}^{1-y}.\end{aligned}$$

/L23

Estimation of Binary Logistic Regression Parameters (Con'd)

Given the independent observations

$$(X_{1,1}, \cdots X_{k,1}, Y_1 = y_1),$$

$$\vdots$$

$$(X_{1,n}, \cdots X_{k,n}, Y_n = y_n),$$

$$\begin{aligned} \mathrm{P}(Y_i = y_i | X_{1,i}, \cdots, X_{k,i}) &= \{\pi(X_{1,i}, \cdots, X_{k,i})\}^{y_i}\{1 - \pi(X_{1,i}, \cdots, X_{k,i})\}^{1-y_i} \\ &= \{p_i(\beta_0, \cdots, \beta_k)\}^{y_i}\{1 - p_i(\beta_0, \cdots, \beta_k)\}^{1-y_i}, \text{ say.} \end{aligned}$$

obs i.

The likelihood is defined by

indpt

$$\begin{aligned} \boxed{\mathcal{L}(\beta_0, \cdots, \beta_k)} &= \mathrm{P}(Y_1 = y_1, \cdots, Y_n = y_n \mid \text{given all } X\text{s}) \\ &= \prod_{i=1}^{n} \mathrm{P}(Y_i = y_i | X_{1,i}, \cdots, X_{k,i}) \\ &= \prod_{i=1}^{n} \{p_i(\beta_0, \cdots, \beta_k)\}^{y_i}\{1 - p_i(\beta_0, \cdots, \beta_k)\}^{1-y_i}, \end{aligned}$$

which is the probability that we observe $Y_1 = y_1, \cdots, Y_n = y_n$ given all Xs.

Estimation of Binary Logistic Regression Parameters (Con'd)

The maximum likelihood estimation (MLE) takes the "logic" that since we observe $Y_1 = y_1, \cdots, Y_n = y_n$, there should be a pretty good chance that the observed outcome happens. Otherwise, we should not observe it.

Hence, the probability that we observe $Y_1 = y_1, \cdots, Y_n = y_n$ given all Xs, namely the likelihood $\mathcal{L}(\beta_0, \cdots, \beta_k)$, should be very large.

We choose MLE $\hat{\beta}_0, \cdots, \hat{\beta}_k$ numerically to maximize the probability $\mathcal{L}(\beta_0, \cdots, \beta_k)$.

Different from the least squares estimation, we do not have a formula for MLE $\hat{\beta}_0, \cdots, \hat{\beta}_k$. The MLE can only be obtained numerically.

Fitted Probabilities

Using MLE $\hat{\beta}_0, \cdots, \hat{\beta}_k$, the estimated mean function is given by:

$$\hat{\mu}\{Y|X\} = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k) \text{ (plug-in idea)}.$$

Fitted

The ~~fitting~~ probabilities are given by

$$\begin{aligned} \hat{\pi}(X) &= \hat{\mu}\{Y|X\} \\ &= g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k) \\ &= \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k}}. \end{aligned}$$

When we talk about fitted probabilities, X is usually from the training dataset (see Lecture Notes 8).

When X_{new} is from the new dataset or the test dataset, we actually talk about prediction.

Prediction of a New Observation

The forecast of probability is given by

$$\begin{aligned}\hat{\pi}(X_{\text{new}}) &= \hat{\mu}\{Y|X_{\text{new}}\} \\ &= g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 X_{1,\text{new}} + \cdots + \hat{\beta}_k X_{k,\text{new}}) \\ &= \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_{1,\text{new}} + \cdots + \hat{\beta}_k X_{k,\text{new}}}}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 X_{1,\text{new}} + \cdots + \hat{\beta}_k X_{k,\text{new}}}}.\end{aligned}$$

Recall that $\mathrm{P}(Y=1|X) = \pi(X)$ and $\mathrm{P}(Y=0|X) = 1 - \mathrm{P}(Y=1|X) = 1 - \pi(X)$.

Thus, if $\mathrm{P}(Y=1|X) > \mathrm{P}(Y=0|X)$ namely $\pi(X) > 0.5$, there is a better chance that $Y=1$ will occur.

Hence, 0.5 is a commonly used threshold for predicting the response.

In conclusion, the prediction for the response Y_{new} at X_{new} is

$$\hat{Y}_{\text{new}} = 1 \text{ if } \hat{\pi}(X_{\text{new}}) > 0.5; \;\; \hat{Y}_{\text{new}} = 0 \text{ otherwise.}$$

Or equivalently, $\hat{Y}_{\text{new}}$ is the category that has the larger forecast of probability.

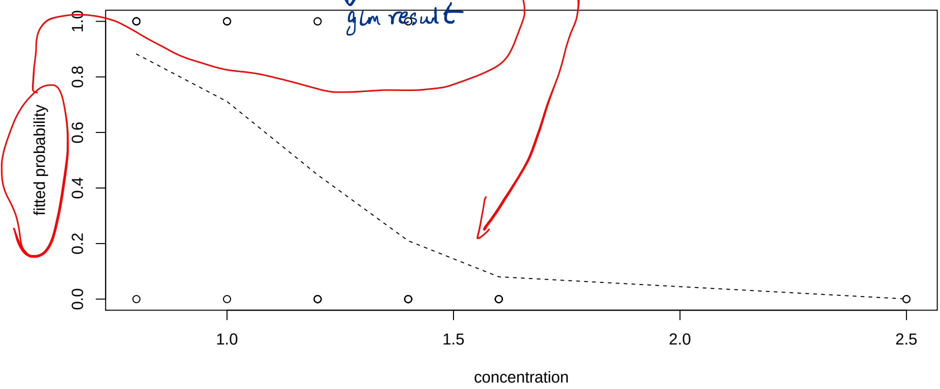
multi-category

Example: Anaesthetic Data (Con'd)

Bernoulli is a special case of

```
#?glm
#fitting the logistic regression
ansth.logit=glm(Response~Concentration,family=binomial(link=logit))
plot(Concentration,Response,xlab="concentration",ylab="fitted probability")
lines(Concentration,ansth.logit$fitted.values,lty=2)
```

glm result




```
detach(a)
```

All the fitted probabilities are between zero and one.

Example: Anaesthetic Data (Con'd)

By using this example, we might be interested in predicting whether a patient will respond to the stimulus if an anaesthetic at a new concentration of 1.5 is given. The forecast of probability is:

```
Xnew=data.frame(Concentration=1.5)
predict(ansth.logit,Xnew,type='response')
```

result of
g(m()

```
##         1
## 0.1322204
```

→ prediction of probability that Y=1

For this patient we predict a response of 0, i.e., we predict that the patient will not respond to the stimulus.