106-1 生物統計學二 實習課

R : Logistic Regression

周芷妤

2017.11.02

d04849010@ntu.edu.tw

大綱

- Review

- Logistic Regression
  - Fit logistic model
  - Test the usefulness of model
  - Add mean response to scatter plot

Review

Review

- Linear regression
  - ✓ Homework 3 & 4
  - ✓ Homework 4 (Lab)

Logistic Regression

Fit logistic model
Test the usefulness of model
Add mean response to scatter plot

Fit generalized linear model

- Fit **logistic** regression model

  model <- **glm**(Y ~ X1 + X2 + … + Xp , family = binomial , data = 資料檔名稱)

  - 告知R現在要建立的是 logistic regression
  - 也可打成family = binomial(link="logit")

- 產生model配適結果的總結

  summary(model)

Example : Low Birth Weight data

Goal: risk factors associated with low infant birth weight
＊資料檔 : lbw.csv (逗號分隔)

Coding Book	
變項名稱	變項描述
low	1 = birth weight of a baby is under 2500g 0 = birth weight of a baby is over 2500g
smoke	smoking status during pregnancy 1 = yes , 0 = no
race	mother's race 1 = white, 2 = black, 3 = other
age	mother's age in years
lwt	mother's weight in pounds at last menstrual period
ptl	number of previous premature labours
ht	1 = history of hypertension 0 = no hypertension
ui	presence of uterine irritability 1 = yes , 0 = no
ftv	number of physician visits in 1st trimester
bwt	birth weight in grams

Example : Low Birth Weight data

Q : lwt及race是否會影響嬰兒出生體重過輕?

Variables

Y : low

X_1: lwt

X_2: race $= \begin{cases} \text{white} \\ \text{black} \\ \text{other} \end{cases}$

Coding book

race	$X_{2(1)}$	$X_{2(2)}$
white	1	0
black	0	1
other	0	0

Reference → other

Model 1

$Y|X \sim \mathrm{Ber}(p_X)$

$p_X = \mathrm{P}(Y = 1|X) = \mathrm{E}[Y|X]$

$$\mathrm{logit}(p_X) = \ln\left(\frac{p_X}{1-p_X}\right) = \beta_0 + \beta_1 X_1 + \beta_{2(1)} X_{2(1)} + \beta_{2(2)} X_{2(2)}$$

$$\Longrightarrow p_X = \frac{e^{\beta_0+\beta_1 X_1+\beta_{2(1)} X_{2(1)}+\beta_{2(2)} X_{2(2)}}}{1+e^{\beta_0+\beta_1 X_1+\beta_{2(1)} X_{2(1)}+\beta_{2(2)} X_{2(2)}}}$$

Meaning of β_1

$$\ln\left(\frac{p_X}{1-p_X}\right) = \beta_0 + \beta_1 X_1 + \beta_{2(1)} X_{2(1)} + \beta_{2(2)} X_{2(2)}$$

$$\Longrightarrow \frac{p_X}{1-p_X} = \frac{\mathrm{P}\left(Y=1 \middle| \boldsymbol{X_1}, X_{2(1)}, X_{2(2)}\right)}{1-\mathrm{P}\left(Y=1 \middle| \boldsymbol{X_1}, X_{2(1)}, X_{2(2)}\right)} = e^{\beta_0+\beta_1 \boldsymbol{X_1}+\beta_{2(1)}X_{2(1)}+\beta_{2(2)}X_{2(2)}}$$

While controlling $X_{2(1)}$ & $X_{2(2)}$,

- $X_1 = x+1 \Longrightarrow \dfrac{\mathrm{P}\left(Y=1 \middle| \boldsymbol{X_1=x+1}, X_{2(1)}, X_{2(2)}\right)}{1-\mathrm{P}\left(Y=1 \middle| \boldsymbol{X_1=x+1}, X_{2(1)}, X_{2(2)}\right)} = e^{\beta_0+\beta_1(x+1)+\beta_{2(1)}X_{2(1)}+\beta_{2(2)}X_{2(2)}}$
- $X_1 = x \Longrightarrow \dfrac{\mathrm{P}\left(Y=1 \middle| \boldsymbol{X_1=x}, X_{2(1)}, X_{2(2)}\right)}{1-\mathrm{P}\left(Y=1 \middle| \boldsymbol{X_1=x}, X_{2(1)}, X_{2(2)}\right)} = e^{\beta_0+\beta_1(x)+\beta_{2(1)}X_{2(1)}+\beta_{2(2)}X_{2(2)}}$

$$\Longrightarrow \text{Odds Ratio (OR) of } X_1 = \frac{e^{\beta_0+\beta_1(x+1)+\beta_{2(1)}X_{2(1)}+\beta_{2(2)}X_{2(2)}}}{e^{\beta_0+\beta_1(x)+\beta_{2(1)}X_{2(1)}+\beta_{2(2)}X_{2(2)}}} = e^{\beta_1}$$

同一種族的人(調整種族的影響後)，lwt 每增加一單位，導致嬰兒出生體重過輕的OR增加e^{β_1}倍

Example : Low Birth Weight data

```
> # 將race轉成dummy variable
> data.lbw$race.w <- ifelse(data.lbw$race=="1", 1, 0)
> data.lbw$race.b <- ifelse(data.lbw$race=="2", 1, 0)

> model.1 <- glm(low ~ lwt + race.w + race.b, family = binomial, data = data.lbw)
> summary(model.1)

Call:
glm(formula = low ~ lwt + race.w + race.b, family = binomial,
    data = data.lbw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3486  -0.8917  -0.7197   1.2527   2.0987

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.283996   0.796553   1.612   0.1070
lwt         -0.015201   0.006438  -2.361   0.0182 *
race.w      -0.481036   0.356646  -1.349   0.1774
race.b       0.599683   0.508863   1.178   0.2386
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 223.28  on 185  degrees of freedom
AIC: 231.28

Number of Fisher Scoring iterations: 4
```

在相同種族下(調整種族的影響後)，
lwt 每增加一單位，導致嬰兒出生體重過輕的危險性(OR)多$e^{-0.015201} = 0.9849141$倍

$\mathrm{H_0}: \beta_1 = 0$ vs. $\mathrm{H_1}: \beta_1 \neq 0$

$Z = \frac{-0.015201}{0.006438} = -2.361$

p-value = $0.0182 < \alpha = 0.05$

Example : Low Birth Weight data

- Test the usefulness of logistic model

$H_0: \beta_1 = \beta_{2(1)} = \beta_{2(2)} = 0$ vs. H_1: at least one $\beta_j \neq 0, \quad j = 1,\ 2(1),\ 2(2)$

建立只有截距項的logistic model

```
> model.0 <- glm(low ~ 1, family = binomial, data = data.lbw)
> anova(model.0, model.1, test="Chisq")
Analysis of Deviance Table

Model 1: low ~ 1
Model 2: low ~ lwt + race.w + race.b
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       188     234.67
2       185     223.28  3   11.395 0.00977 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value = $0.00977 < \alpha = 0.05$

Plot **curve**

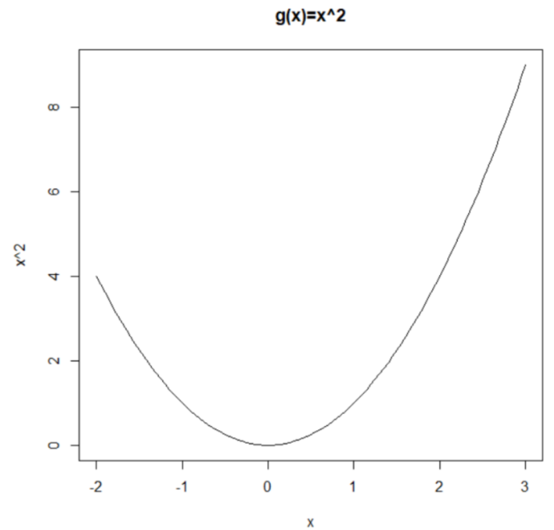
- 繪製曲線

curve(函數名稱 or 表達式, from = a, to = b, add = F)

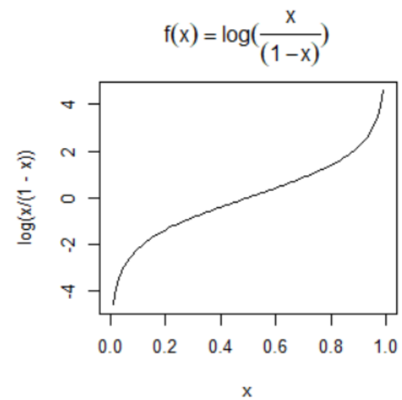
- R內建函數的自變數預設為x
- from = a , to = b → 以x=a到x=b 繪製函數
- add = F → 不在上一張圖加上曲線 (預設為FALSE)

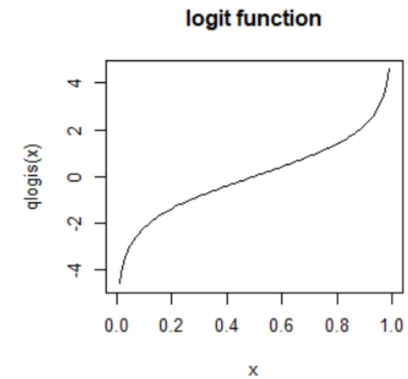
➢ e.g. 繪製 $g(x) = x^2, \quad -2 \leq x \leq 3$

curve(x^2, -2, 3, main = "g(x)=x^2")



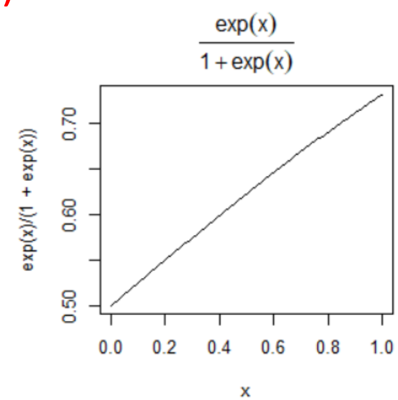
- e.g. 繪製 $f(x) = \text{logit}(x) = \ln(\frac{x}{1-x})$

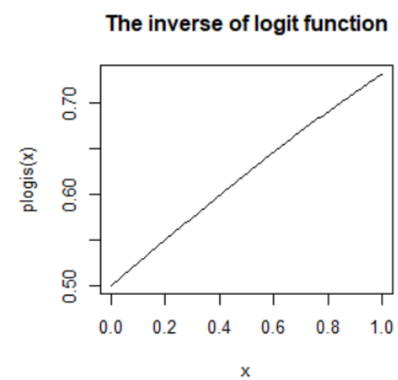




curve(log(x/(1-x)), main = **expression(** f(x)==log(frac(x, (1-x))) **)**)

curve(**qlogis(**x**)**, main = "logit function")





- e.g. 繪製 $\text{logit}^{-1}(x) = (\frac{e^x}{1+e^x})$

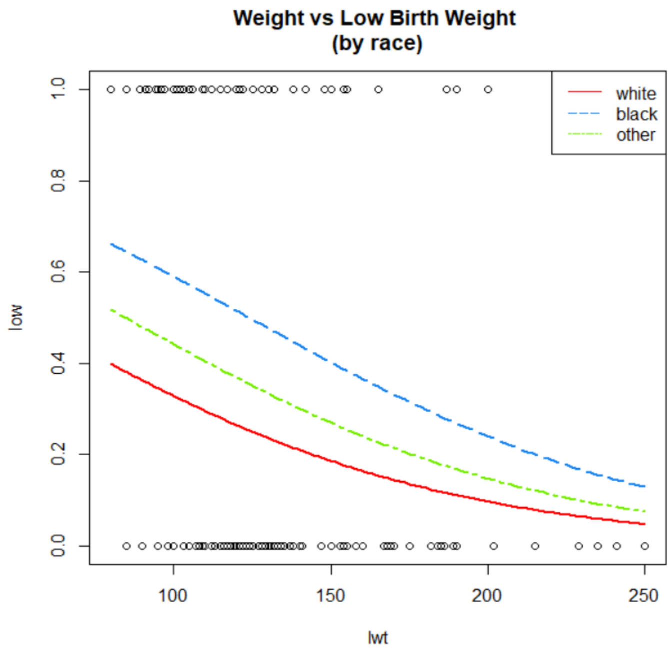
curve(exp(x)/(1+exp(x)), main = **expression(** frac(exp(x),1+exp(x)) **)**)

curve(**plogis(**x**)**, main = "The inverse of logit function")

Add mean response to scatter plot

Method 1 利用predict計算mean response

$$E[Y|\boldsymbol{X_1}, X_{2(1)}, X_{2(2)}] = \frac{e^{\beta_0+\beta_1\boldsymbol{X_1}+\beta_{2(1)}X_{2(1)}+\beta_{2(2)}X_{2(2)}}}{1+e^{\beta_0+\beta_1\boldsymbol{X_1}+\beta_{2(1)}X_{2(1)}+\beta_{2(2)}X_{2(2)}}}$$



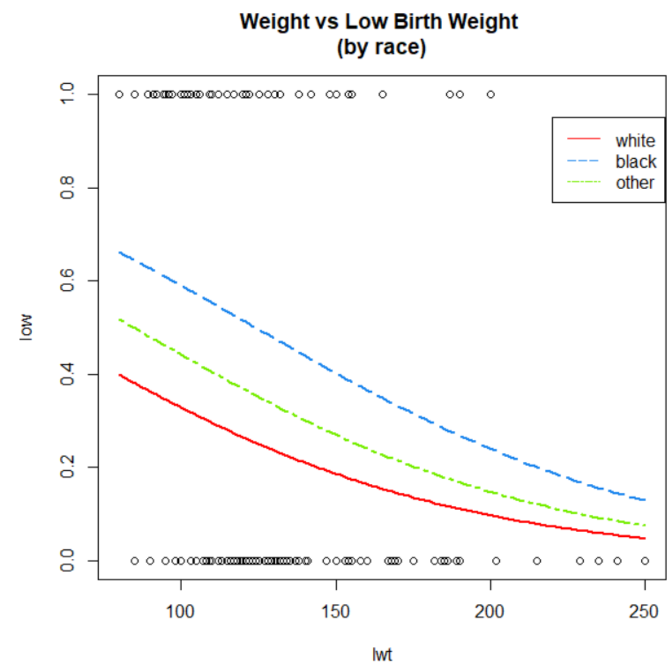

```
> attach(data.lbw)
> plot(lwt, low, main="Weight vs Low Birth Weight \n(by race)")
> # 加上race各類別的 mean response
> curve( predict(model.1, data.frame(lwt=x, race.w=1 , race.b=0), type="response"),
+        add = T, col="red", lty=1, lwd=2)
> curve( predict(model.1, data.frame(lwt=x, race.w=0 , race.b=1), type="response"),
+        add = T, col="dodgerblue2", lty=5, lwd=2)
> curve( predict(model.1, data.frame(lwt=x, race.w=0 , race.b=0), type="response"),
+        add = T, col="chartreuse2", lty=6, lwd=2)
>
> legend("topright", c("white", "black", "other"),
+        col=c("red", "dodgerblue2", "chartreuse2"), lty=c(1, 5, 6))
```

用哪些變項來預測

要預測的類型為 predicted probabilities

Add mean response to scatter plot

Method 2　　利用plogis計算mean response



For race = white,

$$E[Y|\boldsymbol{X_1}, X_{2(1)} = 1, X_{2(2)} = 0] = \frac{e^{\beta_0+\beta_1\boldsymbol{X_1}+\beta_{2(1)}}}{1+e^{\beta_0+\beta_1\boldsymbol{X_1}+\beta_{2(1)}}}$$

$$\Longrightarrow \hat{E}[Y|\boldsymbol{X_1}, X_{2(1)} = 1, X_{2(2)} = 0] = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\boldsymbol{X_1}+\hat{\beta}_{2(1)}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\boldsymbol{X_1}+\hat{\beta}_{2(1)}}}$$

```
> model.1$coefficients
(Intercept)         lwt      race.w      race.b 
 1.28399570 -0.01520088 -0.48103631  0.59968312 
```

```
> # 定義變項
> b0 <- model.1$coefficients[1]
> b1 <- model.1$coefficients[2]
> b21 <- model.1$coefficients[3]
> b22 <- model.1$coefficients[4]
>
> plot(lwt, low, main="Weight vs Low Birth Weight \n(by race)")
> curve(plogis( b0 + b21 + b1*x), add = T, col="red", lty=1, lwd=2)
> curve(plogis( b0 + b22 + b1*x), add = T, col="dodgerblue2", lty=5, lwd=2)
> curve(plogis( b0 + b1*x), add = T, col="chartreuse2", lty=6, lwd=2)
>
> legend(220, 0.95, c("white", "black", "other"),
+        col=c("red", "dodgerblue2", "chartreuse2"), lty=c(1, 5, 6))
```

R內建函數的自變數預設為x

課堂練習

- **Model 2**：在不同種族間懷孕前末次經期的體重(lwt)對於嬰兒出生體重過輕(low)的風險可能不同

  - 請寫下建立的Model 2 (請將符號定義清楚)
  - 請執行 logistic regression，並解釋說明在不同種族之間 lwt 對 low 的影響 (如: 對於白人而言，lwt 每增加一單位…)
  - Model 2有用嗎？
  - 請在scatter plot上畫出不同種族之間 lwt 對 low 的 mean response，並加上legend作說明