Logistic Regression with Categorical Predictors 3/1/22

Simple case: $Y = \begin{cases} 1 & \text{"success"} \\ 0 & \text{"failure"} \end{cases}$ $P(Y=1) = \pi$

Predictors: $X = \begin{cases} 1 \\ 0 \end{cases}$ $Z = \begin{cases} 1 \\ 0 \end{cases}$

$\Rightarrow$ Can display/summarize data in $2 \times 2 \times 2$ table.

Additive model: $\underbrace{\log\left(\frac{\pi}{1-\pi}\right)}$ $= \alpha + \beta_1 x + \beta_2 z$

Interpret coefs:

Covariate pattern:		Logit	
$X=0$	$Z=0$	α	"baseline group"
$X=0$	$Z=1$	$\alpha+\beta_2$	
$X=1$	$Z=0$	$\alpha+\beta_1$	
$X=1$	$Z=1$	$\alpha+\beta_1+\beta_2$	

$\underbrace{}$ exponentiate $\longrightarrow$ odds $\left(\frac{\pi}{1-\pi}\right)$ of each group.

Conditional odds ratios:

If $Z=0$: OR for $x=1$ compared to $x=0$:
(condition)

$$\frac{e^{\alpha+\beta_1}}{e^{\alpha}} = e^{\beta_1}$$

equal

If $Z=1$: OR for $X=1$ compared to $X=0$:

$$e^{\alpha+\beta_1+\beta_2} / e^{\alpha+\beta_2} = e^{\beta_1}$$

Since $\dfrac{\text{odds } X=1 \mid Z}{\text{odds } X=0 \mid Z}$ — same for $Z=0$ & $Z=1$

$\Longleftrightarrow$ no interaction between X & Z

$\Longleftrightarrow$ homogeneous association (3-way table)

Interaction model: $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X + \beta_2 Z + \beta_3 X \cdot Z$

Covariate pattern		Logit
$X=0$	$Z=0$	α
$X=0$	$Z=1$	$\alpha+\beta_2$
$X=1$	$Z=0$	$\alpha+\beta_1$
$X=1$	$Z=1$	$\alpha+\beta_1+\beta_2+\beta_3$

When $Z=0 \Rightarrow$ OR $X=1$ vs. $X=0$: $\dfrac{e^{\alpha+\beta_1}}{e^{\alpha}} = e^{\beta_1}$

$\neq$

When $Z=1 \Rightarrow$ OR $X=1$ vs $X=0$: $\dfrac{e^{\alpha+\beta_1+\beta_2+\beta_3}}{e^{\alpha+\beta_2}} = e^{\beta_1+\beta_3}$

$= e^{\beta_1} \cdot e^{\beta_3}$

$\Rightarrow$ Model allows for non-homogeneous assoc. (interaction)

Interpret β_3? e^{β_3} = Multiplicative effect on the odds ratio of $X=1$ compared to $X=0$ when Z changes from 0 to 1

Categorical predictors w/ more than 2 levels:

K level (categories)

Code K−1 indicator variables

$$\log\left(\frac{\pi}{1-\pi}\right) = \underline{\text{Program}}$$

e.g. Program (Berkeley data) A, B, C, D, E, F

$$= \alpha + \beta_1 B + \beta_2 C + \beta_3 D + \beta_4 E + \beta_5 F$$

$$P = \begin{cases} 1 & \text{Program } P \\ 0 & \text{else} \end{cases}$$

⇒ Program A is baseline.

Note: If not specified, baseline category in R will be 1st category in alphabetical order.

To specify: "relevel" function

Each category has a unique combination of indicator variables:

Variables:	B	C	D	E	F	G
Category A	0	0	0	0	0	0
Category D	0	0	1	0	0	0

Example: (2) Additive: Sex + Program

(1) Interaction: Sex * Program

(2) Fitted model: $\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 0.676 - 0.099 \cdot M$

$$- 0.046B - 1.287C - 1.271D - 1.725E - 3.301F$$

where π = probability of admission,

$$M = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases} \qquad P = \begin{cases} 1 & \text{Program } P \\ 0 & \text{else} \end{cases} \quad P = B, C, D, E, F$$

① Fitted interaction model:

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 1.544 - 1.057M - 0.790B - \cdots - 4.125F$$

$$+ 0.8295\, M\cdot B + 1.182\, M\cdot C + \cdots + 0.8683\, M\cdot F$$

Interpret: Within female applicants $(M=0)$

$$OR \quad \frac{\text{Prog. } C \mid M=0}{\text{Prog. } A \mid M=0}$$

Additive: $$\frac{\exp(0.676 - 1.257)}{\exp(0.676)} = e^{-1.257} = 0.284$$

For female applicants, the estimated odds of admission in Program C are 72% lower than that of program A.

What about $OR \quad \frac{C \mid M=1}{A \mid M=1} \longrightarrow 0.284$ differ in interaction model $\to$ check

(Not in textbook)

Confidence Intervals for (asymptotic) Linear Combinations of Coefficients

General form CI for θ:

$$\hat{\theta} \pm (\text{critical value}) \times SE(\hat{\theta})$$

where the critical value is taken from the asymptotic distribution of $\frac{\hat{\theta}-\theta}{SE(\hat{\theta})}$ or $\frac{\hat{\theta}-\theta}{SD(\hat{\theta})}$

- GM → std. normal

What if we want a CI for $e^{\beta_1+\beta_2}$?

1. CI for $\beta_1+\beta_2$ $\qquad \theta = \beta_1+\beta_2$
2. Exponentiate endpoints.

$$\hat{\theta} = \hat{\beta}_1 + \hat{\beta}_2 \Rightarrow SE(\hat{\theta}) \neq SE(\hat{\beta}_1) + SE(\hat{\beta}_2)$$

since $\hat{\beta}_1$ & $\hat{\beta}_2$ are correlated

$$Cov(\hat{\beta}_1, \hat{\beta}_2) \neq 0$$

Background: Properties of variance:

Two random variables X & Y: $\qquad (\hat{\beta}_1 \text{ \& } \hat{\beta}_2)$

$$Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2ab\, Cov(X,Y)$$

scalars (known constants)

Examples: $Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$

$Var(X-Y) = Var(X) + Var(Y) - 2\,Cov(X,Y)$

<u>Matrix Form:</u>

$$\underset{\text{random vector}}{\underline{X}} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}_{k\times 1}$$

In practice: $\hat{\beta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{k-1} \end{pmatrix}$

$$\underset{k\times 1}{E(\underline{X})} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_k) \end{pmatrix}$$

$$\underset{k\times k}{Var(\underline{X})} = \begin{array}{c|ccccc} & 1 & 2 & 3 & \cdots & k \\ \hline 1 & Var(X_1) & Cov(X_1,X_2) & & \cdots & \\ 2 & Cov(X_2,X_1) & Var(X_2) & & & \\ 3 & Cov(X_3,X_1) & & \ddots & & \\ \vdots & \vdots & & & & \\ k & Cov(X_k,X_1) & & & & Var(X_k) \end{array}$$

$(i,j)^{th}$ cell $= Cov(X_i, X_j)$

Let A be any $m\times k$ matrix.

$$\Rightarrow \underset{m\times 1}{E(A\underline{X})} = A\,E(\underline{X})$$

$$\underbrace{Var(A\underline{X})}_{m\times m} = \underset{m\times k}{A}\ \underset{k\times k}{Var(\underline{X})}\ \underset{k\times m}{A^T}$$

<u>Matrices in R:</u>

To create a matrix:

```
matrix(c(____),
       nrow= ,
       ncol= ,
       byrow = TRUE)
```

Matrix multiplication:

`%*%`

Applied to Logistic Regression:

Ex: $\beta_1 + \beta_2$ — Additive

$$\hat{\beta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_6 \end{pmatrix} \qquad \hat{\beta}_1 + \hat{\beta}_2 = \underbrace{\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}}_{A} \hat{\beta}$$

$\underline{\underline{X}}$

$$SE(\hat{\beta}_1 + \hat{\beta}_2) = \sqrt{A \; vcov(\hat{\beta}) \; A^T}$$