# 106-1 生物統計學二 實習課

## R : Poisson Regression

周芷妤

d04849010@ntu.edu.tw

2017.11.23

# 大綱

- Poisson Regression
  - Fit Poisson model

# Poisson Regression

Fit Poisson model

# Fit generalized linear model

- Fit **Poisson** regression model

model <- **glm**( Y ~ X1 + X2 + … + Xp , offset = log(person-year), family = poisson,
data = 資料檔名稱)

不同樣本之間需以相同的基準比較，
用來校正計數資料(count data)

- 告知R現在要建立的是 poisson regression
- 也可使用 family = poisson(link="log")

(查詢: ?glm.fit、?family)

- model配適結果的總結

**summary**( model )

# Example : Smoke lung data

**Goal**: 每日抽菸量如何影響肺癌死亡率
＊資料檔 : smokelung.csv (逗號分隔)

| Coding Book | |
|---|---|
| 變項名稱 | 變項描述 |
| years.smoke | a factor giving the age that start smoking |
| cigarettes | a factor giving cigarette consumption/day |
| Time | man-years at risk |
| y | number of deaths |

# Example : Smoke lung data

Q:每日抽菸量如何影響肺癌死亡率

**Variables**

$Y$ : y (number of deaths)

$X_1$: years.smoke $=$ { 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59 }

$X_2$: cigarettes $=$ { 0, 1–9, 10–14, 15−19, 20−24, 25−34, 35+ }

**Coding book**

Reference →

| years.smoke | $X_{1(1)}$ | $X_{1(2)}$ | ⋯ | $X_{1(8)}$ |
|---|---|---|---|---|
| 15–19 | 0 | 0 | ⋯ | 0 |
| 20–24 | 1 | 0 | ⋯ | 0 |
| 25–29 | 0 | 1 | ⋯ | 0 |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| 55-59 | 0 | 0 | ⋯ | 1 |

Reference →

| cigarettes | $X_{2(1)}$ | $X_{2(2)}$ | ⋯ | $X_{2(6)}$ |
|---|---|---|---|---|
| 0 | 0 | 0 | ⋯ | 0 |
| 1–9 | 1 | 0 | ⋯ | 0 |
| 10–14 | 0 | 1 | ⋯ | 0 |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| 35+ | 0 | 0 | ⋯ | 1 |

利用**contrasts(變項名稱)**查看

# Example : Smoke lung data

**Model**

offset term

$$\mathrm{E}[Y|X] = m \cdot \lambda_X$$

$$\begin{cases} Y|X \sim \text{Poisson}(m \cdot \lambda_X) \\ \log(\lambda_X) = \beta_0 + \beta_{1(1)}X_{1(1)} + \cdots + \beta_{1(8)}X_{1(8)} + \beta_{2(1)}X_{2(1)} + \cdots + \beta_{2(6)}X_{2(6)} \end{cases}$$

$$\Longrightarrow \quad \lambda_X = e^{\beta_0 + \beta_{1(1)}X_{1(1)} + \cdots + \beta_{1(8)}X_{1(8)} + \beta_{2(1)}X_{2(1)} + \cdots + \beta_{2(6)}X_{2(6)}}$$

| | cigarettes | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | | 1-9 | | 10-14 | | 15-19 | | 20-24 | | 25-34 | | 35+ | |
| years.smoke | y | Time | y | Time | y | Time | y | Time | y | Time | y | Time | y | Time |
| 15-19 | 1 | 10366 | 0 | 3121 | 0 | 3577 | 0 | 4317 | 0 | 5683 | 0 | 3042 | 0 | 670 |
| 20-24 | 0 | 8162 | 0 | 2937 | 1 | 3286 | 0 | 4214 | 1 | 6385 | 1 | 4050 | 0 | 1166 |
| 25-29 | 0 | 5969 | 0 | 2288 | 1 | 2546 | 0 | 3185 | 1 | 5483 | 4 | 4290 | 0 | 1482 |
| 30-34 | 0 | 4496 | 0 | 2015 | 2 | 2219 | 4 | 2560 | 6 | 4687 | 9 | 4268 | 4 | 1580 |
| 35-39 | 0 | 3512 | 1 | 1648 | 0 | 1826 | 0 | 1893 | 5 | 3646 | 9 | 3529 | 6 | 1336 |
| 40-44 | 0 | 2201 | 2 | 1310 | 1 | 1386 | 2 | 1334 | 12 | 2411 | 11 | 2424 | 10 | 924 |
| 45-49 | 0 | 1421 | 0 | 927 | 2 | 988 | 2 | 849 | 9 | 1567 | 10 | 1409 | 7 | 556 |
| 50-54 | 0 | 1121 | 3 | 710 | 4 | 684 | 2 | 470 | 7 | 857 | 5 | 663 | 4 | 255 |
| 55-59 | 2 | 826 | 0 | 606 | 3 | 449 | 5 | 280 | 7 | 416 | 3 | 284 | 1 | 104 |

# Example : Smoke lung data

```
> model.1 <- glm(y~ years.smoke + cigarettes, offset = log(Time), family = poisson, data = smokelung_data )
> summary(model.1)

Call:
glm(formula = y ~ years.smoke + cigarettes, family = poisson,
    data = smokelung_data, offset = log(Time))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8329  -0.8560  -0.3808   0.4241   2.1762

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -12.5784     1.1475 -10.961  < 2e-16 ***
years.smoke20-24   0.9469     1.1548   0.820 0.412202
years.smoke25-29   1.7016     1.0805   1.575 0.115284
years.smoke30-34   3.2029     1.0204   3.139 0.001695 **
years.smoke35-39   3.2423     1.0242   3.166 0.001547 **
years.smoke40-44   4.2088     1.0137   4.152 3.30e-05 ***
years.smoke45-49   4.4476     1.0171   4.373 1.23e-05 ***
years.smoke50-54   4.9048     1.0201   4.808 1.52e-06 ***
years.smoke55-59   5.4134     1.0239   5.287 1.24e-07 ***
cigarettes1-9      1.2200     0.7073   1.725 0.084547 .
cigarettes10-14    2.0991     0.6363   3.299 0.000971 ***
cigarettes15-19    2.3089     0.6327   3.649 0.000263 ***
cigarettes20-24    2.9009     0.5956   4.870 1.11e-06 ***
cigarettes25-34    3.1162     0.5947   5.240 1.61e-07 ***
cigarettes35+      3.6059     0.6048   5.962 2.49e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 445.099  on 62  degrees of freedom
Residual deviance:  51.471  on 48  degrees of freedom
AIC: 201.31

Number of Fisher Scoring iterations: 6
```

在相同years.smoke下 (調整初次抽菸年齡的影響後)，**每日抽35根菸以上者**的肺癌死亡率為**沒有抽菸者**的 $e^{3.6059} = 36.815$ 倍

$H_0: \beta_{2(6)} = 0$ vs. $H_1: \beta_{2(6)} \neq 0$

$Z = \frac{3.6059}{0.6048} = 5.962$

p-value $= 2.49 \times 10^{-9} < \alpha = 0.05$

# Meaning of $\beta_{2(6)}$

$$\log(\lambda_X) = \beta_0 + \beta_{1(1)}X_{1(1)} + \cdots + \beta_{1(8)}X_{1(8)} + \beta_{2(1)}X_{2(1)} + \cdots + \beta_{2(6)}X_{2(6)}$$

$$\Longrightarrow \quad \lambda_X = e^{\beta_0+\beta_{1(1)}X_{1(1)}+\cdots+\beta_{1(8)}X_{1(8)}+\beta_{2(1)}X_{2(1)}+\cdots+\beta_{2(6)}X_{2(6)}}$$

**While controlling $X_{1(1)}, \ldots, X_{1(8)}$ & $X_{2(1)}, \ldots, X_{2(5)}$,**

- $X_{2(6)} = 1$

  $\Longrightarrow \mathrm{E}[Y|X_{1(1)}, \ldots, X_{1(8)}, X_{2(1)}, \ldots, X_{2(5)}, X_{2(6)} = 1] = m \cdot e^{\beta_0+\beta_{1(1)}X_{1(1)}+\cdots+\beta_{1(8)}X_{1(8)}+\beta_{2(1)}X_{2(1)}+\cdots+\beta_{2(6)}\times(1)}$

- $X_{2(6)} = 0$

  $\Longrightarrow \mathrm{E}[Y|X_{1(1)}, \ldots, X_{1(8)}, X_{2(1)}, \ldots, X_{2(5)}, X_{2(6)} = 0] = m \cdot e^{\beta_0+\beta_{1(1)}X_{1(1)}+\cdots+\beta_{1(8)}X_{1(8)}+\beta_{2(1)}X_{2(1)}+\cdots+\beta_{2(6)}\times(0)}$

$$\Longrightarrow \text{Rate Ratio (RR) of } X_{2(6)} = \frac{m \cdot e^{\beta_0+\beta_{1(1)}X_{1(1)}+\cdots+\beta_{2(6)}\times(1)}}{m \cdot e^{\beta_0+\beta_{1(1)}X_{1(1)}+\cdots+\beta_{2(6)}\times(0)}} = e^{\beta_{2(6)}}$$

調整其他變項的影響後，
$X_{2(6)} = 1$ 的RR為 $X_{2(6)} = 0$ 的 $e^{\beta_{2(6)}}$ 倍

# Homework

**example4：**抽菸習慣與死亡率的關係

＊請先匯入資料檔(需附上code)： example4.csv (逗號分隔)

- 利用R進行 Poisson regression，判斷調整年齡組別變項後，抽菸習慣和死亡率的關係：
  - 請寫下建立的迴歸模型 (定義清楚符號代表的意思)
  - 請問抽菸習慣是否會影響死亡率?(請說明根據什麼結果判斷)
  - 校正年齡組別後，Smoke的Rate Ratio為?
  - 對於有抽菸習慣且年齡大於55歲的人，估計的死亡率為? ($\hat{\lambda}_X$ =?)

| Coding Book | |
|---|---|
| 變項名稱 | 變項描述 |
| Agegroup | 年齡組別<br>(1:<25歲/ 2:25-35歲 /3:35-45歲/ 4:45-55歲/ 5:>55歲) |
| Smoke | 有無抽菸習慣(0: 沒有抽菸; 1:有抽菸習慣) |
| Death | 死亡數 |
| py | 觀察到的人年(person-year) |