Lecture 10: Logistic regression

BTBI30081

統計應用方法Applied Methods in Statistics

2025/4/23

Odds ratio

	豪宅		
車位	是 (1)	不是 (0)	total
有 (1)	146 (*a*)	3609 (*b*)	3755
無 (0)	391 (*c*)	6173 (*d*)	6564
total	537	9783	10319

- 有附車位的房子會是豪宅的勝算 (odds)：$\frac{146}{3609}$
- 沒附車位的房子會是豪宅的勝算 (odds)：$\frac{391}{6173}$
- 有車位對上沒有車位的豪宅勝算比 (odds ratio)：$\left(\frac{146}{3609}\right)/\left(\frac{391}{6173}\right)$

(RMD_example 10.2)

Logistic regression

- The respond variable Y is binary (e.g., yes or no, success or failure).
- $$\log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_P x_P$$

  Y: response variable (binary) (random variable),

  $x_1, \cdots, x_P$: covariates (continuous or binary) (known values),

  $\alpha_0, \alpha_1, \cdots, \alpha_P$: regression coefficients (unknown parameters).

Interpretation of regression coefficients

- $\log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_P x_P$

α_0 = the log odds $\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right)$ of $x_1 = \cdots = x_P = 0$

α_p = the log odds ratio for every 1 unit increase in x_p when holding other covariates unchanged

Example

- $\log\left(\frac{\Pr(豪宅=1)}{\Pr(豪宅=0)}\right) =$
  $-3.05 - 0.68\ (車位) + 0.56\ (有無管理組織),$
- $\exp(\alpha_0) =$ 0.05 = 對那些沒附車位且也沒有管理組織的房子，他們會是豪宅的勝算 (odds)
- $\exp(\alpha_1) =$ 0.51 = 對管理組織相同的房子，有車位對上沒有車位的豪宅勝算比 (odds ratio)
- $\exp(\alpha_2) =$ 1.75 =對車位狀態相同的房子，有管理組織對上沒有管理組織的豪宅勝算比 (odds ratio)

(RMD_example 10.3)

Parameter estimation: the maximum likelihood method

- Maximum likelihood is based on choosing the values of regression coefficient α's that make the probability of observing your result as large as possible.
- Regression coefficient β's in linear regression can also be obtained by maximum likelihood.

How good the logistic regression is

- In linear regression, the coefficient of determination R^2, which represents the fraction of the total variation of the data explained by the used model, can is used to measure how good the model is.
- In logistic regression, R^2 is not a valid goodness-of-fit measurement; need to develop a quantity in logistic regression.

Deviance

- Deviance =
$$2 \times \log\left(\frac{\text{probability of observing your result} \mid \text{data}}{\text{probability of observing your result} \mid \text{model}}\right)$$
- The smaller the deviance, the closer your model to the data (good fit).

Logistic regression vs. linear regression

- Significant tests for Ho : $\alpha_p = 0$
- Polynomial regression
- Dummy variables
- Interaction
- Confounding

Logistic regression results

```
Call:
glm(formula = 豪宅 ~ 車位 + 有無管理組織, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3993  -0.3993  -0.3047  -0.2870   2.7388

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.04690    0.07861 -38.760  < 2e-16 ***
車位          -0.67973    0.10620  -6.400 1.55e-10 ***
有無管理組織   0.55761    0.10181   5.477 4.33e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4220.0  on 10318  degrees of freedom
Residual deviance: 4167.6  on 10316  degrees of freedom
AIC: 4173.6

Number of Fisher Scoring iterations: 6
```

$\hat{\alpha}_0$

$\hat{\alpha}_1$

$\hat{\alpha}_2$

Logistic regression results

```
Call:
glm(formula = 豪宅 ~ 車位 + 有無管理組織, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3993  -0.3993  -0.3047  -0.2870   2.7388

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.04690    0.07861 -38.760  < 2e-16 ***
車位         -0.67973    0.10620  -6.400 1.55e-10 ***
有無管理組織  0.55761    0.10181   5.477 4.33e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4220.0  on 10318  degrees of freedom
Residual deviance: 4167.6  on 10316  degrees of freedom
AIC: 4173.6

Number of Fisher Scoring iterations: 6
```

$SE(\hat{\alpha}_0)$

$SE(\hat{\alpha}_1)$

$SE(\hat{\alpha}_2)$

Logistic regression results

p−value for Ho: $\alpha_0 = 0$

```
Call:
glm(formula = 豪宅 ~ 車位 + 有無管理組織, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3993  -0.3993  -0.3047  -0.2870   2.7388

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.04690    0.07861 -38.760  < 2e-16 ***
車位         -0.67973    0.10620  -6.400 1.55e-10 ***
有無管理組織  0.55761    0.10181   5.477 4.33e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4220.0  on 10318  degrees of freedom
Residual deviance: 4167.6  on 10316  degrees of freedom
AIC: 4173.6

Number of Fisher Scoring iterations: 6
```

p−value for Ho: $\alpha_1 = 0$

p−value for Ho: $\alpha_2 = 0$

Logistic regression results

```
Call:
glm(formula = 豪宅 ~ 車位 + 有無管理組織, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3993  -0.3993  -0.3047  -0.2870   2.7388

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.04690    0.07861 -38.760  < 2e-16 ***
車位         -0.67973    0.10620  -6.400 1.55e-10 ***
有無管理組織  0.55761    0.10181   5.477 4.33e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4220.0  on 10318  degrees of freedom
Residual deviance: 4167.6  on 10316  degrees of freedom
AIC: 4173.6

Number of Fisher Scoring iterations: 6
```

deviance