Analysis of Categorical Data
Chapter 5 and 6: Logistic Regression

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. make inference for a logistic model,
2. perform model diagnostic/selection,
3. estimate odds ratio from a logistic model,
4. test conditional independence,
5. test homogeneous association.

Logistic Regression

In general, a logistic regression model is of the form

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\beta}^T \boldsymbol{x}_i, \quad i = 1, ..., n,$$

where $\pi_i = P(Y_i = 1 \mid \boldsymbol{x}_i)$ and $Y_i \mid \boldsymbol{x}_i \sim \text{Binomial}(m_i, \pi_i)$.

- The link function is the logit link $g(\pi) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$.
- We can fit the model by IRLS.

Interpretation of Logistic Model

Consider the logistic model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Then,

$$\begin{aligned} & \frac{P\left(Y_i = 1 | x_{i1} = a+1, x_{i2}, ..., x_{ip}\right) / P\left(Y_i = 0 | x_{i1} = a+1, x_{i2}, ..., x_{ip}\right)}{P\left(Y_i = 1 | x_{i1} = a, x_{i2}, ..., x_{ip}\right) / P\left(Y_i = 0 | x_{i1} = a, x_{i2}, ..., x_{ip}\right)} \\ = \ & \frac{\exp\left\{\beta_0 + \beta_1 (a+1) + \beta_2 x_2\right\}}{\exp\left\{\beta_0 + \beta_1 a + \beta_2 x_2\right\}} = \exp\left\{\beta_1\right\} \end{aligned}$$

is the (conditional) odds ratio, adjusting for other covariates.

Generally speaking, β_j is the expected change in the log odds for one unit increase in x_{ij}, holding the other terms fixed.

Sampling: Prospective or Retrospective

Sometimes (e.g., a case-control study), X is random instead of Y.

Prospective study

	Cancer		
Smoking	Yes	No	Total
Yes	n_{11}	n_{12}	n_{1+}
No	n_{21}	n_{22}	n_{2+}

Case-Control study

	Cancer	
Smoking	Yes	No
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}
Total	n_{+1}	n_{+2}

In a prospective study, we have $P(Y \mid X)$. In a case-control study, we have $P(X \mid Y)$. We can still build a logistic model to model $P(Y \mid X)$ among the selected subjects in a case-control study, the β can still be estimated. Hence, we can still estimate the odds ratio.

Qualitative Explanatory Variables

In our course, we mainly work with logistic models with categorical x.

Consider an $I \times 2$ table. In row i, let y_i be the number of successes out of n_i trials. We can treat y_i as $Y_i \sim \text{Bin}\left(n_i, \pi_i\right)$.

- The corresponding logit model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i,$$

  expressed as the model in one-way ANOVA.

- Using dummy variables, the model becomes

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_{I-1} x_{I-1} + \beta_I x_I,$$

  where x_j's are the dummy variables.

Identification and Interpretation

For identification, we need $\beta_1 = 0$ or $\beta_I = 0$, or other conditions.

- Suppose that $\beta_I = 0$ such that $x_i = i$ if the observations are in row i. Then,

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha + \beta_i, \quad i = 1, ..., I-1, \\ \log\left(\frac{\pi_I}{1-\pi_I}\right) &= \alpha. \end{aligned}$$

- α is the log odds for row I, and $\alpha + \beta_i$ is the log odds for row i.
- β_i is the log odds ratio between row i and I.
- $\beta_i - \beta_j$ is the log odds ratio between row i and j.

Test β_i

We know from the general theory of GLM that

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\right).$$

- We can test individual β_i using

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\left[\left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\right]_{ii}}} \sim N(0,\ 1).$$

- We can test a linear combination $\boldsymbol{c}^T\boldsymbol{\beta}$ using

$$\frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{c}^T\left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{c}}} \sim N(0,\ 1).$$

Saturated Model and Null Model

Consider the model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i,$$

where $i = 1, 2, ..., I$.

- The model is saturated if the model has I parameters. The MLE satisfies $\hat{\pi}_i = Y_i / n_i$.
- In the null model $\beta_i = 0$ for all i, then $\text{logit}\,(\pi_i) = \alpha$ and

$$P\,(Y = 1 \mid X = i) = \frac{\exp\{\alpha\}}{1 + \exp\{\alpha\}},$$

  implying that X and Y are independent.
  - What can we use null deviance for?

Ordinal Predictor

If we formulate the model as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i,$$

then we treat the categorical X as nominal.

If X is ordinal, then it is always difficult to treat. Two alternatives are

1. Assign scores and use scores as the continuous covariates. But the scores can affect the results.
2. Treat ordinal X as nominal X. But we have information loss.

Example: Heart Disease

We have a sample of males. The response variable is whether they developed coronary heart disease. The explanatory variable is the blood pressure level.

```
Data

##   with without pressure
## 1    3     153     <117
## 2   17     235  117-126
## 3   12     272  127-136
## 4   16     255  137-146
## 5   12     127  147-156
## 6    8      77  157-166
## 7   16      83  167-186
## 8    8      35     >186
```

Pressure as Ordinal: Model Fitting

If we treat pressure as an ordinal variable, then we can assign scores of your choice to it and fit a logistic model.

```
Data$score <- c(111.5, 121.5, 131.5, 141.5, 151.5, 161.5, 176.5, 191.5)
Logit <- glm(cbind(with, without) ~ score, family = binomial, data = Data)
summary(Logit)


##
## Call:
## glm(formula = cbind(with, without) ~ score, family = binomial,
##     data = Data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0617  -0.5977  -0.2245   0.2140   1.8501
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.082033   0.724320  -8.397  < 2e-16 ***
## score        0.024338   0.004843   5.025 5.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.0226  on 7  degrees of freedom
## Residual deviance:  5.9092  on 6  degrees of freedom
## AIC: 42.61
##
## Number of Fisher Scoring iterations: 4
```

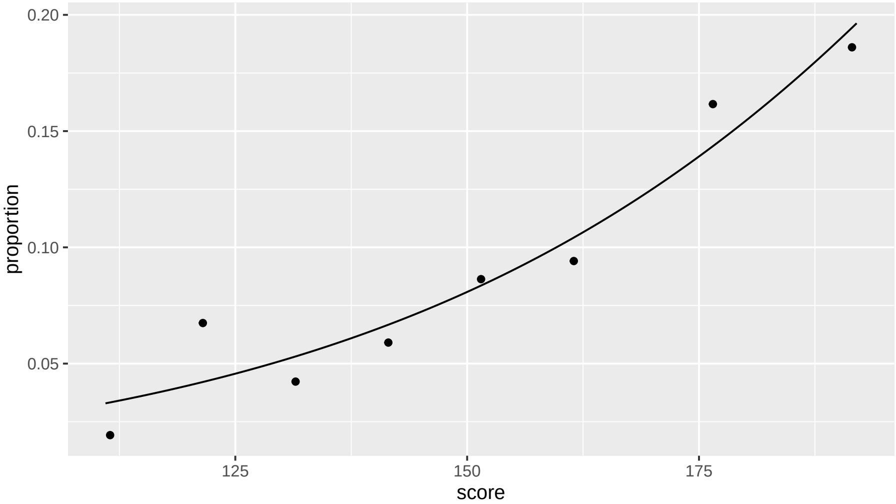
Pressure as Ordinal: Residuals

We can compute the Pearson residual and the standardized Pearson residual. The latter is closer to $N(0,1)$ if the model holds.

```
## Pearson residual
residuals(Logit, type = "pearson")

##          1          2          3          4          5          6          7
## -0.9794311  2.0057103 -0.8133348 -0.5067270  0.1175833 -0.3042459  0.5134721
##          8
## -0.1394648

## Standardized Pearson residual
rstandard(Logit, type = "pearson")

##          1          2          3          4          5          6          7
## -1.1057850  2.3746058 -0.9452701 -0.5727440  0.1260886 -0.3260730  0.6519547
##          8
## -0.1773473
```

Pressure as Ordinal: Plots

We can plot the observed proportions and compare them with the fitted curve.

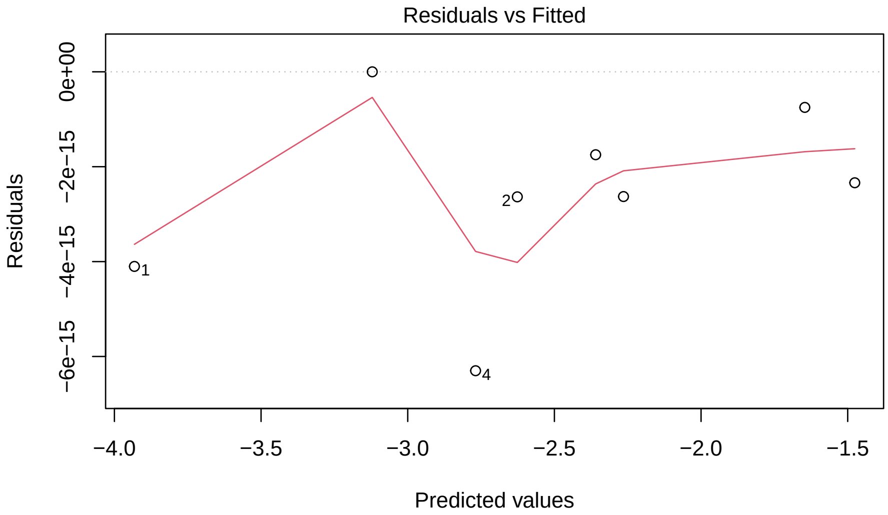


Pressure as Nominal: Model Fitting

If we treat pressure as an nominal variable, then the model fits the data perfectly.

```
##
## Call:
## glm(formula = cbind(with, without) ~ pressure, family = binomial,
##     data = Data)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.9318     0.5830  -6.744 1.54e-11 ***
## pressure>186      2.4559     0.7025   3.496 0.000472 ***
## pressure117-126   1.3055     0.6348   2.057 0.039731 *
## pressure127-136   0.8109     0.6534   1.241 0.214543
## pressure137-146   1.1632     0.6374   1.825 0.068030 .
## pressure147-156   1.5725     0.6566   2.395 0.016615 *
## pressure157-166   1.6675     0.6913   2.412 0.015858 *
## pressure167-186   2.2856     0.6438   3.550 0.000385 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3.0023e+01  on 7  degrees of freedom
## Residual deviance: 3.2196e-14  on 0  degrees of freedom
## AIC: 48.701
##
```

Pressure as Nominal: Zero Residuals



Multiway Table

- The model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i$$

  can be used for a two-way table of size $I \times 2$.
- If we have a three-way table of size $I \times 2 \times K$, then we can consider the model

$$\begin{aligned} \log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) &= \alpha + \beta_i^X + \beta_k^Z, \\ \text{or} \quad \log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) &= \alpha + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}, \end{aligned}$$

  where π_{ik} is the success probability when $X = i$ and $Z = k$.

Homogeneous Association and Conditional Independence

Consider the model

$$\log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = \alpha + \beta_i^X + \beta_k^Z,$$

for a $2 \times 2 \times K$ model. At a fixed level of $Z = k$, the log odds ratio is

$$\left(\alpha + \beta_1^X + \beta_k^Z\right) - \left(\alpha + \beta_0^X + \beta_k^Z\right) = \beta_1^X - \beta_0^X = \beta_1^X,$$

if the identification restriction is $\beta_0^X = 0$.

- The conditional odds ratio is $\exp\left(\beta_1^X\right)$ for any $Z = k$, which means that the $2 \times 2 \times K$ table has homogeneous XY association.
- If we further have $\beta_1^X = 0$, then the conditional odds ratio is 1 and $X \perp Y \mid Z$ (conditional independence).

Logit Model to Test Conditional Independence

In a $2 \times 2 \times K$ table, the logistic model becomes

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad k = 1, ..., K,$$

where $x_i = 0$ or 1. Testing conditional independence $X \perp Y \mid Z$ is equivalent to testing $H_0 : \beta = 0$ in the model.

1. If we assume homogeneous XY association, then
   - the Wald test statistic is $\hat{\beta}/\text{SE}$.
   - the likelihood ratio test compares the deviances between the model with $\beta = 0$ and the model with estimated β.
2. More generally, we can compare the model

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k^Z, \quad k = 1, ..., K.$$

   and the saturated model using the deviance as a goodness-of-fit test of the model.

Test Conditional Independence: Example

A clinical trial

		Response	
Study	Treatment	Success	Failure
1	Drug	11	25
	Placebo	10	27
2	Drug	16	4
	Placebo	22	10
3	Drug	14	5
	Placebo	7	12
4	Drug	2	14
	Placebo	1	16

Cochran-Mantel-Haenszel Test

The Cochran-Mantel-Haenszel test is a non-model-based test of conditional independence in a $2 \times 2 \times K$ table.

- When $K = 1$, regardless of sampling, under the independence assumption, conditioning on both sets of marginal totals, the only free cell is n_{11} that follows the hypergeometric distribution

$$P(n_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1} - t}}{\binom{n_{++}}{n_{+1}}}.$$

  (Fisher's exact test).
- The mean and variance of the hypogeometric distribution are

$$\begin{aligned} \mathbb{E}(n_{11}) &= \frac{n_{1+}n_{+1}}{n_{++}}, \\ \mathrm{var}(n_{11}) &= \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n_{++}^2(n_{++} - 1)}. \end{aligned}$$

Partial Table

When $K > 1$, in each partial table k, we conditional on the row margins and column margins. When the conditional independence assumption holds, then n_{11k} follows a hypergeometric distribution with

$$\begin{aligned} \mu_{11k} = \mathbb{E}\left(n_{11k}\right) &= \frac{n_{1+k}n_{+1k}}{n_{++k}}, \\ \operatorname{var}\left(n_{11k}\right) &= \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2\left(n_{++k}-1\right)}. \end{aligned}$$

The Cochran-Mantel-Haenszel test statistic is

$$\text{CMH} = \frac{\left[\sum_k \left(n_{11k} - \mu_{11k}\right)\right]^2}{\sum_k \operatorname{var}\left(n_{11k}\right)},$$

which has a large-sample chi-squared null distribution with degree of freedom 1.

Common Odds Ratio

In the logit model

$$\operatorname{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad k = 1, ..., K,$$

the conditional odds ratio is $\exp(\beta)$ for any $Z = k$ (homogeneous association). The ML estimate of the common odds ratio is $\exp\left(\hat{\beta}\right)$, where $\hat{\beta}$ is the MLE of β.

The Mantel-Haenszel estimator is

$$\hat{\theta}_{MH} = \frac{\sum_k \left(n_{11k} n_{22k} / n_{++k}\right)}{\sum_k \left(n_{12k} n_{21k} / n_{++k}\right)}.$$

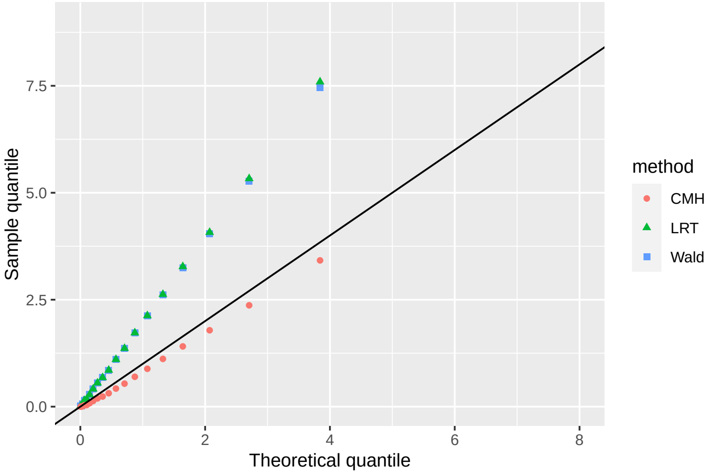
More on Cochran-Mantel-Haenszel

The CMH test can also work well when $K \to \infty$ as $n \to \infty$ (sparse table).

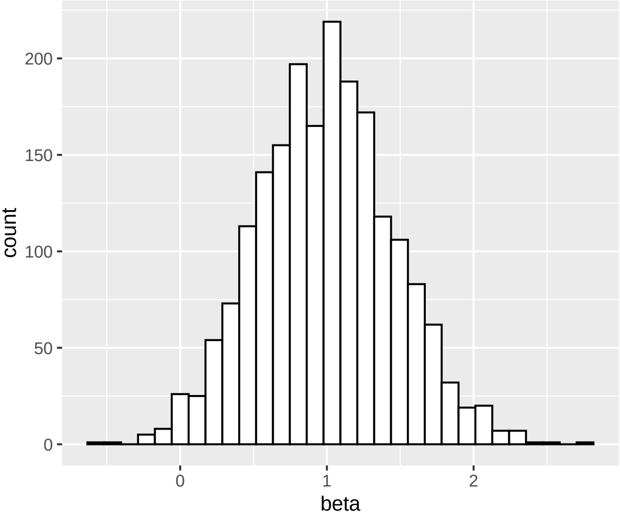
- This occurs for example for paired data: for each k, the treatment is offered only to one subject, and the control is offered only to one subject.
- In this case, $n = 2K$ and the number of observations in each partial table is 2.
- If a logistic model is fitted, the number of parameters is $1 + 1 + (K - 1)$.

The MH estimator of common odds ratio is generally preferred over the ML estimator if K is large and the tables are sparse.

What's The Problem? Simulation When $\beta = 0$



What's The Problem? Simulation of MLE When $\beta = 0.5$



Meta-Analysis

Suppose that we have K studies for the same research question. Each study yields a 2×2 table. We can combine information from all studies and refer analysis to the $2 \times 2 \times K$ table.

		Response	
Study	Treatment	Success	Failure
1	Drug	11	25
	Placebo	10	27
2	Drug	16	4
	Placebo	22	10
3	Drug	14	5
	Placebo	7	12
4	Drug	2	14
	Placebo	1	16

Conditional Association

Suppose that we have (X, Y, Z) in a $2 \times 2 \times K$ table, where Z is a control variable. Let $\{\mu_{ijk}\}$ be the cell expected frequencies corresponding to $(X = i, Y = j, Z = k)$. Then,

$$\text{conditional odds ratio: } \theta_{XY(k)} = \frac{\mu_{11k}/\mu_{12k}}{\mu_{21k}/\mu_{22k}}, \text{ fixing } Z = k,$$

$$\text{marginal odds ratio: } \theta_{XY} = \frac{\mu_{11+}/\mu_{12+}}{\mu_{21+}/\mu_{22+}},$$

are generally not the same, where $\mu_{ij+} = \sum_k \mu_{ijk}$.

However, they will be the same if

1. either Z and X are conditionally independent,
2. or Z and Y are conditionally independent.

These conditions are called the collapsibility conditions.

Back to Logit Models

Consider a $2 \times 2 \times K$ table. The logit model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad k = 1, ..., K,$$

has the same treatment effect β for each $Z = k$.

- The XY conditional odds ratio is $\exp(\beta)$.
- The marginal odds ratio θ_{XY} can be different from $\exp(\beta)$, since we do not have the collapsibility conditions.

Consider a $2 \times 2 \times K$ table. The logit model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i,$$

satisfies the collapsibility condition $Y \perp Z \mid X$. Hence, the XY conditional odds ratio $\exp(\beta)$ is the same as the marginal odds ratio.

Test Homogeneous Association

The logit model

$$\operatorname{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad k = 1, ..., K,$$

has homogeneous association. Hence, we can test the goodness-of-fit of the model as a tool to test homogeneous association.

		Response	
Study	Treatment	Success	Failure
1	Drug	11	25
	Placebo	10	27
2	Drug	16	4
	Placebo	22	10
3	Drug	14	5
	Placebo	7	12
4	Drug	2	14
	Placebo	1	16