Logistic regression

$$d = \begin{cases} 1: & \text{case} \\ 0: & \text{control} \end{cases}$$

$x = (x_0, x_1, \ldots, x_p)$ vector of covariates $(x_0 \equiv 1)$

Logistic regression model:

$$P(d = 1 \mid x) = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_j\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_j\right)}$$

or

$$\ln \frac{P(d = 1 \mid x)}{1 - P(d = 1 \mid x)} = \textstyle\sum_{j=0}^{p} \beta_j x_j$$

Logistic regression likelihood

Sample of n individuals

d_i : case status of individual i

$x_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$: vector of covariates of individual i

Likelihood function:

$$L(\beta \mid d, x) = \prod_{i=1}^{n} \frac{\exp\left(d_i \sum_{j=0}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}$$

Coding of genotypes

Diallelic marker locus $\{A, a\}$

Genotype coding:

Genotype	x_2	x_1
AA	1	0
Aa	0	1
aa	0	0

Allele coding:

Genotype	x_1
AA	2
Aa	1
aa	0

Logistic regression and genotype coding

D_i : number of cases with i copies of allele A

C_i : number of controls with i copies of allele A (c.f. CC/7)

Logistic regression likelihood function:

$$
\begin{aligned}
&L(\beta_0, \beta_1, \beta_2 \mid D_2, D_1, D_0, C_2, C_1, C_0) \\
&= \left(\frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\beta_0 + \beta_2)}\right)^{D_2} \cdot \left(\frac{1}{1 + \exp(\beta_0 + \beta_2)}\right)^{C_2} \\
&\quad \cdot \left(\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}\right)^{D_1} \cdot \left(\frac{1}{1 + \exp(\beta_0 + \beta_1)}\right)^{C_1} \\
&\quad \cdot \left(\frac{\exp(\beta_0)}{1 + \exp(\beta_0)}\right)^{D_0} \cdot \left(\frac{1}{1 + \exp(\beta_0)}\right)^{C_0}
\end{aligned}
$$

Logistic regression and genotype coding

The logistic regression likelihood function is the product of three terms of the form

$$s^u \cdot (1-s)^v,$$

which takes its maximum value at $s = u/(u+v)$, i.e., if $\widehat{\beta}_i$ denotes the maximum likelihood estimate of β_i, then

$$\frac{\exp(\widehat{\beta}_0)}{1+\exp(\widehat{\beta}_0)} = \frac{D_0}{D_0 + C_0} \Rightarrow \exp(\widehat{\beta}_0) = \frac{D_0}{C_0}$$

$$\frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1)}{1+\exp(\widehat{\beta}_0 + \widehat{\beta}_1)} = \frac{D_1}{D_1 + C_1} \Rightarrow \exp(\widehat{\beta}_0 + \widehat{\beta}_1) = \frac{D_1}{C_1} \Rightarrow \exp(\widehat{\beta}_1) = \frac{D_1 \cdot C_0}{C_1 \cdot D_0}$$

$$\frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_2)}{1+\exp(\widehat{\beta}_0 + \widehat{\beta}_2)} = \frac{D_2}{D_2 + C_2} \Rightarrow \exp(\widehat{\beta}_0 + \widehat{\beta}_2) = \frac{D_2}{C_2} \Rightarrow \exp(\widehat{\beta}_2) = \frac{D_2 \cdot C_0}{C_2 \cdot D_0}$$

Score test

Assume that the parameter θ is decomposed into $\theta = (\psi, \eta)$ and the hypothesis of interest is

$$H_0 : \psi = \psi_0$$

$l(\theta)$: log-likelihood function

Score function:

$$U(\theta) = \begin{pmatrix} \frac{\partial l(\theta)}{\partial \psi} \\ \frac{\partial l(\theta)}{\partial \eta} \end{pmatrix} = \begin{pmatrix} U_\psi(\theta) \\ U_\eta(\theta) \end{pmatrix}$$

observed Fisher information:

$$i_n(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} = \begin{pmatrix} i_{\psi\psi}(\theta) & i_{\psi\eta}(\theta) \\ i_{\eta\psi}(\theta) & i_{\eta\eta}(\theta) \end{pmatrix}$$

Score test

inverse of the observed Fisher information:

$$i_n^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi}(\theta) & i^{\psi\eta}(\theta) \\ i^{\eta\psi}(\theta) & i^{\eta\eta}(\theta) \end{pmatrix}$$

Let $\tilde{\theta} = (\psi_0, \tilde{\eta})$ denote the ML estimate under H_0.

Then, the Score test statistic for $H_0 : \psi = \psi_0$ is given by

$$U_\psi^T(\tilde{\theta}) \cdot i^{\psi\psi}(\tilde{\theta}) \cdot U_\psi(\tilde{\theta}),$$

which is asymptotically χ_d^2 distributed (d: dimension of ψ) under H_0.

Logistic regression and Armitage's trend test

Exercise:

Show that in a logistic regression model for a single diallelic marker locus with allele coding of genotypes, the Score test for $H_0 : \beta_1 = 0$ is equivalent to Armitage's trend test.

Conditional logistic regression: Motivation

Assume a 1:1 matching of cases and controls and consider one case-control pair. According to the logistic regression model, the probability that individual k ($k = 1, 2$) of this pair is a case is

$$p_k = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{kj}\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{kj}\right)}$$

and the probability that individual k is a control is

$$p_{2+k} = \frac{1}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{kj}\right)}$$

Conditional logistic regression: Motivation

Therefore, the conditional probability that individual 1 is a case and individual 2 is a control, given that exactly one of the two individuals is a case, becomes

$$\frac{p_1 \cdot p_4}{p_1 \cdot p_4 + p_2 \cdot p_3} = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{1j}\right)}{\exp\left(\sum_{j=0}^{p} \beta_j x_{1j}\right) + \exp\left(\sum_{j=0}^{p} \beta_j x_{2j}\right)}$$

$$= \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{1j}\right)}{\exp\left(\sum_{j=1}^{p} \beta_j x_{1j}\right) + \exp\left(\sum_{j=1}^{p} \beta_j x_{2j}\right)}$$

Conditional logistic regression

H strata

n_h individuals in stratum h ($1 \leq h \leq H$)

first m_h individuals in stratum h are cases and the remaining $n_h - m_h$ individuals are controls (i.e., m_h : $(n_h - m_h)$ matching in stratum h)

Conditional logistic regression likelihood function:

$$L(\beta \mid x) = \prod_{h=1}^{H} \frac{\prod_{i=1}^{m_h} \text{exp}\left(\sum_{j=1}^{p} \beta_j x_{hij}\right)}{\sum \prod_{k=k_1}^{k_{m_h}} \text{exp}\left(\sum_{j=1}^{p} \beta_j x_{hkj}\right)},$$

where the summation is over all $\binom{n_h}{m_h}$ subsets $\{k_1, \ldots, k_{m_h}\}$ of m_h individuals chosen from n_k individuals in stratum h.

Conditional logistic regression and the TDT

Parental genotypes	Offspring genotypes	Genotypes of pseudo-controls	likelihood contribution	number of families
AA, Aa	AA	Aa, Aa, AA	$\frac{\exp(\beta_1)}{2 + 2\exp(\beta_1)}$	n_1
Aa, aa	Aa	aa, aa, Aa		
AA, Aa	Aa	Aa, AA, AA	$\frac{1}{2 + 2\exp(\beta_1)}$	n_2
Aa, aa	aa	aa, Aa, Aa		
Aa, Aa	AA	Aa, Aa, aa	$\frac{\exp(2\beta_1)}{(1 + \exp(\beta_1))^2}$	n_3
Aa, Aa	Aa	AA, Aa, aa	$\frac{\exp(\beta_1)}{(1 + \exp(\beta_1))^2}$	n_4
Aa, Aa	aa	AA, Aa, Aa	$\frac{1}{(1 + \exp(\beta_1))^2}$	n_5
AA, AA	AA	AA, AA, AA	$\frac{1}{4}$	
AA, aa	Aa	Aa, Aa, Aa		
aa, aa	aa	aa, aa, aa		

Conditional logistic regression and the TDT

Score test for $H_0 : \beta_1 = 0$:

$$\begin{aligned}
\ln L(\beta_1) &= (n_1 + 2n_3 + n_4)\beta_1 \\
&\quad -(n_1 + n_2 + 2n_3 + 2n_4 + 2n_5)\ln(1+\exp(\beta_1)) \\
&\quad -(n_1 + n_2)\ln(2) \\
\frac{\partial \ln L}{\partial \beta_1} &= (n_1 + 2n_3 + n_4) - (n_1 + n_2 + 2n_3 + 2n_4 + 2n_5)\frac{\exp(\beta_1)}{1+\exp(\beta_1)}
\end{aligned}$$

$\Rightarrow$

$$U_{\beta_1}(0) = (n_1 + 2n_3 - n_2 - 2n_5)/2$$

Conditional logistic regression and the TDT

$$\begin{aligned}\frac{\partial^2 \ln L}{\partial \beta_1^2} &= -(n_1+n_2+2n_3+2n_4+2n_5)\frac{\exp(\beta_1)}{(1+\exp(\beta_1))^2}\\ i_n(0) &= (n_1+n_2+2n_3+2n_4+2n_5)/4\end{aligned}$$

⇒ Score test statistic is

$$U_{\beta_1}(0)\cdot i_n^{-1}(0)\cdot U_{\beta_1}(0) = \frac{(n_1+2n_3-n_2-2n_5)^2}{(n_1+n_2+2n_3+2n_4+2n_5)}$$

$b = n_1 + 2n_3 + n_4$ is the number of heterozygous parents who transmitted allele A and $c = n_2 + n_4 + 2n_5$ is the number of heterozygous parents who transmitted allele a. Therefore, the Score test statistic can be written as $(b-c)^2/(b+c)$.

Interaction between two diallelic marker loci

u_1, u_2: genotype coding of first marker locus $\{A, a\}$

v_1, v_2: genotype coding of second marker locus $\{B, b\}$

logistic regression model:

$$\begin{aligned}\ln \frac{P(d=1 \mid x)}{1-P(d=1 \mid x)} &= \delta_0 + \alpha_1 u_1 + \alpha_2 u_2 + \beta_1 v_1 + \beta_2 v_2 \\ & \gamma_{11} u_1 v_1 + \gamma_{12} u_1 v_2 + \gamma_{21} u_2 v_1 + \gamma_{22} u_2 v_2\end{aligned}$$

Hypotheses of interest:

$$H_0 : \alpha_1 = \alpha_2 = \beta_1 = \beta_2 = \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$$

(no main and no interaction effect)

$$H_0^I : \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$$

(no interaction effect)

Interaction between two diallelic marker loci

For $i, j \in \{0, 1, 2\}$, let f_{ij} denote the penetrance of the two-locus genotype with i copies of allele A and j copies of allele B.

Example: $f_{10} = P(\text{affected} \mid Aa, bb)$

It can be shown that H_0^I is true if and only if the two-locus penetrances can be factorized, i.e.,

$$f_{ij} = s_i \cdot t_j$$

for all $(0 \leq i, j \leq 2)$ and appropriately chosen $(s_i)_{i=0,1,2}$ and $(t_j)_{j=0,1,2}$.

Interaction between two diallelic marker loci

Example 1: Two-locus penetrances for the REZ-REZ model

	BB	Bb	bb
AA	1	0	0
Aa	0	0	0
aa	0	0	0

Example 2: Two-locus penetrances for the heterogeneity model

	BB	Bb	bb
AA	1	1	1
Aa	1	0	0
aa	1	0	0