outline

1. Multiple Logistic Regression
2. Extensions to Logistic Regression
3. Sample Size Calculation in Logistic Regression Model

Multiple Logistic Regression

13.8

Introduction

- we learned about the Mantel-Haenszel test and the Mantel Extension test, which are techniques for controlling for a single categorical covariate *C* while assessing the association between a dichotomous disease variable *D* and a categorical exposure variable *E*. But if

  a) E is continuous
  b) or *C* is continuous
  c) or there are several confounding variables $C_1, C_2, \ldots,$ each of which may be either categorical or continuous.

- then it is either difficult or impossible to use the preceding methods to control for confounding.

Example

Infectious Disease

- *Chlamydia trachomatis*（沙眼衣原体） is a microorganism that has been established as an important cause of non-gonococcal urethritis, pelvic inflammatory disease, and other infectious diseases.

- A study of risk factors for *C. trachomatis* was conducted in a population of 431 female college students.

- Because multiple risk factors may be involved, several risk factors must be controlled for simultaneously in analyzing variables associated with *C. trachomatis*.

General model

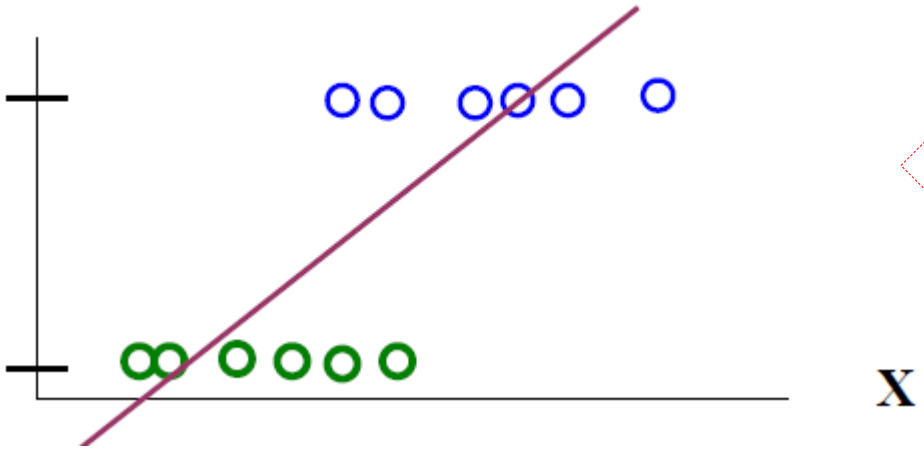
- Regression model might be considered:

$$p = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

- where p =probability of disease.

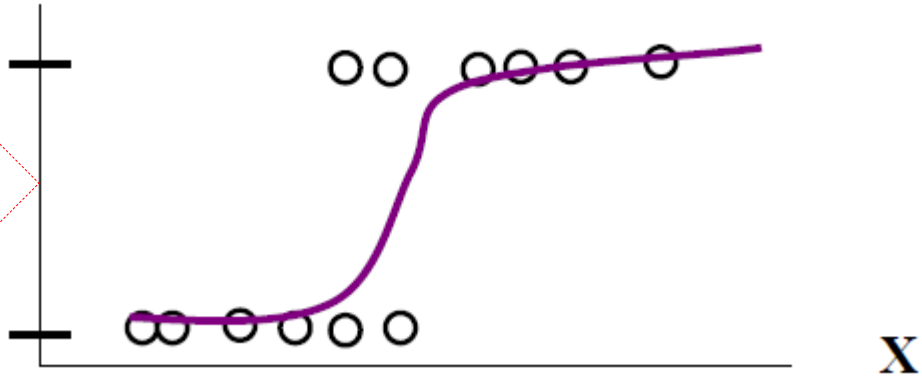
- But the right-hand side could be less than 0 or greater than 1 for certain values of $x_1, \ldots, x_k$, predicted probabilities that are either less than 0 or greater than 1 could be obtained, which is impossible.
- So, take the logit (logistic) transformation of p as follows:

$$\text{logit}(p) = \ln \frac{p}{1-p}$$



a bad-looking fitted line

by using the logistic function (or other functions like probit, complementary log-log function,…), the fit can be largely improved



Multiple Logistic Regression

- $x_1, \ldots, x_k$ are a collection of independent variables
- y is a binomial-outcome variable with probability of success $= p$,

- then the multiple logistic-regression model is given by

$$\text{logit}(p) = \ln \frac{p}{1-p} = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

- or, equivalently,

$$p = \frac{e^{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

Interpretation of Regression Parameters

Categorical variable

- If we refer to the independent variables as exposure variables, individual A is exposed and individual B is not exposed.

	Independent variable							
Individual	1	2	…	$j-1$	j	$j+1$	…	k
A	x_1	x_2	…	x_{j-1}	1	x_{j+1}	…	x_k
B	x_1	x_2	…	x_{j-1}	0	x_{j+1}	…	x_k

- The logit of the probability of success for individuals A and B are given by

$$\text{logit}(p_A) = \alpha + \beta_1 x_1 + \cdots \beta_{j-1} x_{j-1} + \beta_j (1) + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k$$
$$\text{logit}(p_B) = \alpha + \beta_1 x_1 + \cdots \beta_{j-1} x_{j-1} + \beta_j (0) + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k$$

Contd.

- So

$$\text{logit}(p_A) - \text{logit}(p_B) = \beta_j$$

- That is,

$$\ln \frac{p_A/(1-p_A)}{p_B/(1-p_B)} = \beta_j$$

Or

$$\frac{p_A/(1-p_A)}{p_B/(1-p_B)} = e^{\beta_j}$$

- Remember that A is exposed and individual B is not exposed, we can rewrite it as follows:

$$OR = \frac{Odds_A}{Odds_B} = e^{\beta_j}$$

Estimation of *ORs* in Multiple Logistic Regression

- x_j: dichotomous exposure variable,
  - present: coded as 1
  - absent: coded as 0

- The *OR* relating this exposure variable to the dependent variable is estimated by

$$\widehat{OR} = e^{\widehat{\beta}_j}$$

- This relationship expresses the disease–exposure OR after controlling for all other variables in the logistic-regression model.
- a two-sided 100% ×(1 − α) CI for the true *OR* is given by

$$\left(e^{\widehat{\beta}_j - z_{1-\alpha/2} se(\widehat{\beta}_j)}, \qquad e^{\widehat{\beta}_j + z_{1-\alpha/2} se(\widehat{\beta}_j)}\right)$$

Infectious Disease
Revisited

Risk factor	Regression coefficient $(\widehat{\beta}_j)$	Standard error $se(\widehat{\beta}_j)$	z $\widehat{\beta}_j/se(\widehat{\beta}_j)$
Constant	-1.637		
Black race	$+2.242$	0.529	$+4.24$
Lifetime number of sexual partners	$+0.102$	0.040	$+2.55$
Among users of non-barrier[a]			
Methods of contraception[b]			

$$\widehat{OR} = e^{+2.242} = 9.4$$

a 95% CI for *OR* is given by

$$\left(e^{2.242-1.96(0.529)}, e^{2.242+1.96(0.529)}\right) = (3.3, 26.5) \not\ni 1$$

Contingency Table Analysis and logistic model

We can estimate the *OR* relating *D* to *E* in either of two equivalent ways:

		E(exposure)	
		$+$	$-$
D (disease)	$+$	a	b
	$-$	c	d

- directly from the 2 $\times$ 2 table: $OR = \frac{ad}{bc}$
- set up a logistic-regression model,

Let p_E = probability of disease D *occurs* given exposure status E, where

$p_0 = P(D|E = 0) = \exp(\gamma_0) / \{1+\exp(\gamma_0)\}$and

$p_1 = P(D|E = 1) = \exp(\gamma_1) / \{1 + \exp(\gamma_1)\}$.

Denote $\gamma_0 = \alpha$ and $\gamma_1 = \alpha + \beta$, then

$$\ln\{p(D|E)/(1 - p(D|E))\} = \alpha + \beta E$$

$$OR = \frac{p_1/(1 - p_1)}{p_0/(1 - p_0)} = \frac{\exp(\gamma_1)}{\exp(\gamma_0)} = e^{\beta}$$

Contd.

1. For prospective or cross-sectional studies, we can estimate the probability of disease among exposed (p_E) and unexposed ($p_{\bar{E}}$) as follows:
   - From the 2 $\times$ 2 table:

$$p_E = a/(a+c), p_{\bar{E}} = b/(b+d)$$

   - From the logistic-regression model:

$$p_E = e^{\hat{\alpha}+\hat{\beta}} / \left(1 + e^{\hat{\alpha}+\hat{\beta}}\right), p_{\bar{E}} = e^{\hat{\alpha}} / \left(1 + e^{\hat{\alpha}}\right)$$

2. For case–control studies, it is impossible to estimate absolute probabilities of disease unless the sampling fraction of cases and controls from the reference population is known, which is almost always *not* the case.

Revisited (Chap. 10)
Breast Cancer

- Suppose we are interested in the association between the incidence of breast cancer and the age at first childbirth.

- Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan. case

- Controls were chosen from women of comparable age who were in the hospital at the same time as the cases but who did *not* have breast cancer.

- These women are divided into two categories:
  - women whose age at first birth was ≤29 years
  - women whose age at first birth was ≥30 years

Age at first birth			
Status	≥ 30	≤ 29	Total
Case	683	2537	3220
Control	1498	8747	10245
Total	2181	11284	13465

Estimate the OR directly from 2×2 table

```
> dat <- matrix(c(683,1498,2537,8747), nrow=2,
+ dimnames=list(c("D+","D"),c(">=30","<=29")))
> oddsratio(dat, log=F)
[1] 1.571982
```

Or you can set up a logistic-regression model

```
> # 重构原始数据
> d <- c(rep(1,683+2537), rep(0,1498+8747))
> e <- c(rep(1,683), rep(0,2537), rep(1,1498),
rep(0,8747))
> dat.glm <- glm(d ~ e, family = "binomial")
> exp(coef(dat.glm)[2]) # odds ratio
       e
1.571982
```

Continuous variable

- Similarly, if we refer to the independent variables as exposure variables, individual A is exposed and individual B is not exposed.

	Independent variable							
Individual	1	2	…	$j-1$	j	$j+1$	…	k
A	x_1	x_2	…	x_{j-1}	$x_j+\Delta$	x_{j+1}	…	x_k
B	x_1	x_2	…	x_{j-1}	x_j	x_{j+1}	…	x_k

- The logit of the probability of success for individuals A and B are given by

$$\text{logit}(p_A) = \alpha + \beta_1 x_1 + \cdots \beta_{j-1} x_{j-1} + \beta_j \left(x_j + \Delta\right) + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k$$
$$\text{logit}(p_B) = \alpha + \beta_1 x_1 + \cdots \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \cdots + \beta_k x_k$$

Estimation of *ORs* in Multiple Logistic Regression

- x_j: a continuous independent variable, two individuals:
  - $(x_1, \dots, x_{j-1}, x_j + \Delta, \mathrm{x_{j+1}}, \dots, x_k)$
  - $(x_1, \dots, x_{j-1}, \ x_j, \ \mathrm{x_{j+1}}, \dots, x_k)$

- The *OR* relating this exposure variable to the dependent variable is estimated by

$$\widehat{OR} = e^{\widehat{\beta}_j \Delta}$$

- after controlling for all other variables

- a two-sided 100% $\times$(1 − α) CI for the true *OR* is given by

$$\left(e^{\widehat{\beta}_j \Delta - z_{1-\alpha/2} se(\widehat{\beta}_j)\Delta}, \qquad e^{\widehat{\beta}_j \Delta + z_{1-\alpha/2} se(\widehat{\beta}_j)\Delta}\right)$$

Infectious Disease Revisited

Risk factor	Regression coefficient $\left(\widehat{\beta}_j\right)$	Standard error $se\left(\widehat{\beta}_j\right)$	z $\widehat{\beta}_j / se\left(\widehat{\beta}_j\right)$
Constant	-1.637		
Black race	$+2.242$	0.529	$+4.24$
Lifetime number of sexual partners	$+0.102$	0.040	$+2.55$
Among users of non-barrier[a]			
Methods of contraception[b]			

Let $\Delta = 1$

$$\widehat{OR} = e^{+0.102\times 1} = 1.11$$

a 95% CI for *OR* is given by

$$\left(e^{0.102-1.96(0.040)}, e^{0.102+1.96(0.040)}\right) = (1.02, 1.20) \not\ni 1$$

Hypothesis Testing

- How can the statistical significance of the risk factors be evaluated?

- To test the hypothesis

$$H_0: \beta_j = 0, \text{all other } \beta_i \neq 0, \text{ vs. } H_1: \text{all } \beta_j \neq 0$$

- The test statistic

$$z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1) \text{ under } H_0$$

- should only be used if there are at least 20 successes and 20 failures, respectively, in the data set.

Infectious Disease
Revisited

Risk factor	Regression coefficient $(\widehat{\beta}_j)$	Standard error $se(\widehat{\beta}_j)$	z $\widehat{\beta}_j/se(\widehat{\beta}_j)$
Constant	-1.637		
Black race	$+2.242$	0.529	$+4.24$
Lifetime number of sexual partners	$+0.102$	0.040	$+2.55$
Among users of non-barrier[a] Methods of contraception[b]			

The *p*-values are given by

$$P(\text{race}) = 1 \times \left(1 - \Phi(4.24)\right) < 0.001$$

$$P(\text{number of sexual partners}) = 2 \times \left(1 - \Phi(2.55)\right) = 0.011$$

Thus both variables are significantly associated with *C. trachomatis.*

Prediction

- wish to predict the probability of disease (*p*) for a subject with covariate values $x_1, \ldots, x_k$

- compute the linear predictor

$$\hat{L} = \hat{\alpha} + \hat{\beta}_1 x_1 + \cdots \hat{\beta}_k x_k$$

- The point estimate of p

$$\hat{p} = \frac{e^{\hat{\mathrm{L}}}}{1 + \mathrm{e}^{\hat{\mathrm{L}}}}$$

Assessing GOF: Residuals

1. Pearson residuals:

$$r_i = \frac{y_i - \hat{p}_i}{se(\hat{p}_i)}$$

where

- $y_i = I(\text{the } i\text{th observation is a success})$ <= ungrouped data
- y_i = proportion of successes among the *i*th group of observations <= grouped data

2. Deviance residuals:

$$r_i = sign(y_i - \hat{p}_i)\sqrt{-2[y_i \log \hat{p}_i + (1 - y_i)\log(1 - \hat{p}_i)\,]}$$

- You can get the residuals using the function residuals() in R

Extensions to Logistic Regression

1. Matched/Conditional Logistic Regression
2. Polychotomous Logistic Regression
3. Ordinal Logistic Regression

Matched example: Cancer

- A nested case–control design
  - Case: 235 women with breast cancer occurring between 1990 and 2000.
  - Control: One or two were selected per case, yielding a total of 346.
  - The controls were matched on age, time of day of blood draw, fasting status of blood draw, and previous use of post-menopausal hormones.

- The matched sets (case and 1 or 2 controls) were analyzed at the same time for a plasma estradiol.

- How should the association between plasma estradiol and breast cancer be assessed?

Conditional Logistic Regression

- wish to assess the association between the incidence of breast cancer (*D*) and plasma estradiol (*x*)

- but wish to control for other covariates ($z_1, z_2, \ldots, z_k$), denoted in summary by z.
  - age, parity (i.e., number of children), family history of breast cancer, and others.

- Subdivide the data into *S* matched sets ($i = 1, \ldots, S$).
  - a single case and n_i controls, where $n_i \geq 1$ and n_i may vary among matched sets.

a logistic model

$$\text{logit}\left(Pr\left(D_{ij} = 1\right)\right)$$
$$= \alpha_i I(\text{in the } i\text{th matched set}) + \beta x_{ij} + \gamma z_{ij}$$

- nuisance parameters that we do not want to estimate
- we cannot determine α_i because the matched sets are small

Conditional Logistic Regression

Conditional Logistic Regression

- the conditional probability that the jth member of a matched set is a case given that there is exactly one case in the matched set, denoted by p_{ij}.

$$p_{ij} = Pr(D_{ij} = 1 \mid \sum_{k=1}^{n_i} D_{ik} = 1) = \frac{Pr(D_{ij} = 1) \prod_{k=1, k \neq j}^{n_i} Pr(D_{ik} = 0)}{\sum_{l=1}^{n_i} Pr(D_{il} = 1) \prod_{k \neq l}^{n_i} Pr(D_{ik} = 0)}$$

$$= \frac{\exp(\alpha_i + \beta x_{ij} + \underset{\sim}{\gamma}\, \underset{\sim}{z}_{ij}) / \prod_{k=1}^{n_i} [1 + \exp(\alpha_i + \beta x_{ik} + \underset{\sim}{\gamma}\, \underset{\sim}{z}_{ik})]}{\sum_{l=1}^{n_i} \exp(\alpha_i + \beta x_{il} + \underset{\sim}{\gamma}\, \underset{\sim}{z}_{il}) / \prod_{k=1}^{n_i} [1 + \exp(\alpha_i + \beta x_{ik} + \underset{\sim}{\gamma}\, \underset{\sim}{z}_{ik})]}$$

$$= \exp(\beta x_{ij} + \underset{\sim}{\gamma}\, \underset{\sim}{z}_{ij}) / \sum_{l=1}^{n_i} \exp(\beta x_{il} + \underset{\sim}{\gamma}\, \underset{\sim}{z}_{il})$$

Interpretation of Parameters

- Use maximum likelihood methods to find estimates of β and γ which maximize $L = \prod_{i=1}^{S} p_{ij_i}$, where j_i = case in the ith matched set.

- Two subjects in the $i$the matched set
  - A case(j) and a control(l):

$$x_{ij} = x_{il} + 1$$

- The relative risk that the subject with the higher exposure is the case,

$$RR = Pr\left(D_{ij} = 1\right)/Pr(D_{il} = 1) = \exp(\beta)$$

- R function: clogit {Survival}

Polychotomous Logistic Regression(PLR)

- In some cases, we have a categorical outcome variable with more than two categories.
- Often we might have a single control group to be compared with multiple case groups
- or a single case group to be compared with multiple control groups.
- Also known as Multinomial Logistic Regression

Example
Cancer

- Breast cancers are commonly typed using a biochemical assay to determine estrogen receptor (ER) and progesterone receptor (PR) status.

- Tumors can be jointly classified as
  - ER positive (ER+) vs. ER negative (ER−) status
  - PR positive (PR+) vs. PR negative (PR−) status.

- A study was performed to determine risk factor profiles for specific types of breast cancer according to ER/PR status .

- There were 2096 incident cases of breast cancer from 1980−2000, of which 1281 were ER+/PR+, 417 were ER−/PR−, 318 were ER+/PR−, and 80 were ER−/PR+. There was a common control group for all types of breast cancers.

PLR model

$$Pr(\text{1st outcome category}) = \frac{1}{1+\sum_{r=2}^{Q}\exp\left(\alpha_r + \sum_{k=1}^{K}\beta_{rk}x_k\right)}$$

$$Pr(q\text{th outcome category}) = \frac{\exp\left(\alpha_q + \sum_{k=1}^{K}\beta_{qk}x_k\right)}{1+\sum_{r=2}^{Q}\exp\left(\alpha_r + \sum_{k=1}^{K}\beta_{rk}x_k\right)},\ q = 2,\ldots,Q$$

$$\Rightarrow\ odds_{q,S}$$

$$= \frac{\Pr(\text{subject S is in the qth outcome category})}{\Pr\big(\text{subject S is in the 1st outcome category(control group)}\big)}$$

$$= \exp\left[\alpha_r + \sum_{k=1}^{K}\beta_{rk}x_k\right]$$

- R function : mlogit in “mlogit” package

Contd.

- two individuals:
  - A: $(x_1, \dots, x_{j-1}, x_j + 1, \mathrm{x_{j+1}}, \dots, x_k)$
  - B: $(x_1, \dots, x_{j-1},\ x_j,\ \mathrm{x_{j+1}}, \dots, x_k)$

1. the *OR* for being in category q vs. category 1 for subject A vs. subject B

$$\frac{odds_{q,A}}{odds_{q,B}} = \exp(\beta_{qk}) \equiv OR_{qk}$$

2. In general, the *OR* for being in outcome category $q1$ vs. outcome category $q2$ for subject A compared with subject B

$$\frac{odds_{q_1,A}}{odds_{q_2,B}} = \exp(\beta_{q_1,k} - \beta_{q_2,k})$$

Revisited

Cancer

Group	Beta	*se*	*p*-value	RR[b] (95% CI)	Number of cases
no breast cancer	(ref)			1.0	
ER+/PR+	0.00029	0.00009	0.001	1.12 (1.04–1.20)	1281
ER+/PR–	0.00022	0.00017	0.20	1.09 (0.96–1.24)	318
ER–/PR+	0.00015	0.00037	0.68	1.06 (0.80–1.40)	80
ER–/PR–	–0.00003	0.00017	0.86	0.99 (0.87–1.12)	417

[a]Cumulative grams of alcohol before menopause (g/day × years).

[b]The relative risk for 1 drink per day of alcohol from age 18 to age 50 ≅ 12 grams alcohol/drink × 32 years = 384 gram-years × Beta after controlling for 21 other breast cancer risk factors.

The RR for 1g-years/day $= \exp[-0.00003 \times 384] = 0.99$

Ordinal Logistic Regression

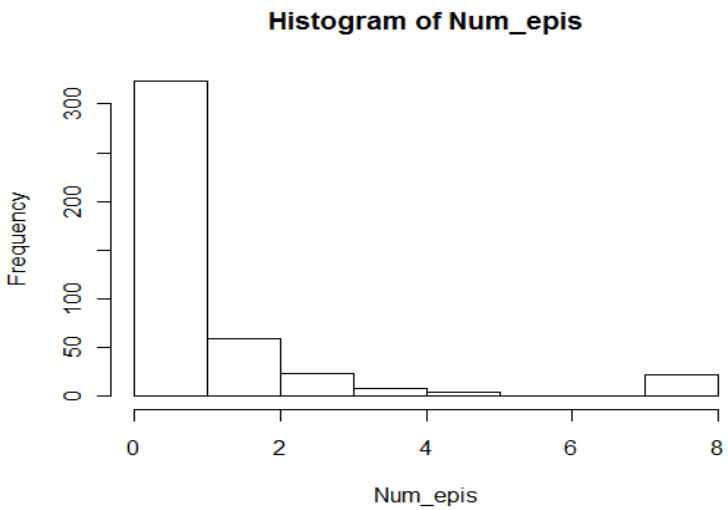
- We first looked at logit estimation in the context of a binary dependent var.
- Then we looked at in a matched set
  - Using Conditional logit
- Then we added the possibility of 3 or more unordered categories for the dependent var.
  - You estimate these using Polychotomous /Multinomial logit
- Now we'll turn to the case of 3 or more ordered categories
  - Educational Attainment: < High School → High School → College → Graduate Degree

Example
Sports Medicine

- an observational study among about 400 members of several tennis clubs in the Boston area.

- The objective: examine risk factors for tennis elbow.

- Prediction: Numbers of current or previous episodes of tennis elbow they had.
  - The distribution ranged from 0 to 8 and was very skewed.
  - elected to categorize the number
  - of episodes into 3 categories (0/1/2+).



Ordinal Logistic Regression

- Outcome variable (y): has c ordered categories $(c \geq 2)$
- Covariates $(x_1, \dots, x_k)$

- Then an ordinal logistic regression model is defined by

$$\log \frac{\Pr(y \leq j)}{\Pr(y \geq j+1)} = \alpha_j - \beta_1 x_1 - \cdots - \beta_k x_k, j = 1, \dots, c-1.$$

$$\Rightarrow e^{\beta_q} = \left(\text{odds that } y \leq j \middle| x_q = x\right) / \left(\text{odds that } y \leq j \middle| x_q = x\right)$$
$$\equiv \text{odds ratio for } y \leq j \text{ given } x_q = x \text{ vs. } x_q = x - 1$$
$$\text{holding all oterh variables constant}$$

the same for each value of j.
$\Rightarrow$ also named cumulative odds or proportional odds ordinal logistic regression model

Revisited

Sports Medicine

R funtion: polr { MASS) or lrm {Design}

					odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Const(1)	-2.9064	0.6013	-4.83	0.0000			
Const(2)	-4.4591	0.6231	-7.16	0.0000			
Age	0.0592	0.1095	5.40	0.0000	1.06	1.039	1.084
Sex	0.3945	0.1852	2.13	0.0331	1.48	1.034	2.137
material__current							
2	0.3177	0.2273	1.40	0.1621	1.37	0.881	2.150
3	0.5537	0.2426	2.28	0.0225	1.73	1.083	2.806

older players and females are likely to have more episodes of tennis elbow

contd.

					odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Const(1)	-2.9064	0.6013	-4.83	0.0000			
Const(2)	-4.4591	0.6231	-7.16	0.0000			
Age	0.0592	0.1095	5.40	0.0000	1.06	1.039	1.084
Sex	0.3945	0.1852	2.13	0.0331	1.48	1.034	2.137
material__current							
2	0.3177	0.2273	1.40	0.1621	1.37	0.881	2.150
3	0.5537	0.2426	2.28	0.0225	1.73	1.083	2.806

In general, users of wood racquets have the least number of episodes of tennis elbow, and users of composite racquets have the greatest; users of metal racquets are in between.

Sample size calculation

in logistic regression model

Sample size calculation in logistic regression model

- Logistic model

$$\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right)=\beta_0+\beta_1 x$$

- Hypothesis test

$$H_0\colon\ \beta_1=0 \text{ vs. } H_1\colon\ \beta_1\neq 0$$

- The rationale is if Y is not related to X, then $\mu_1=\mu_2$, where $\mu_1=E(X|Y=0)$ and $\mu_2=E(X|Y=1)$

- Then the hypothesis testing on β_1 is equivalent to

$$H_0\colon\ \mu_1=\mu_2 \text{ vs. } H_1\colon \mu_1\neq\mu_2$$

- We know that, the sample size being required for testing the equality of two independent sample means μ_1 and μ_2 is,

$$n = \sigma^2 \left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \left[\frac{(k+1)^2}{k}\right] / (\mu_1 - \mu_2)^2,$$

Where σ^2 is the common variance of two normal distributions to be compared, and $k = n_2/n_1$.

- In order to use this formula, we need to establish the relationship between $(k, \mu_1, \mu_2, \sigma)$ and β_1

- $k = \frac{n_2}{n_1} = \frac{P(Y=1|x=EX)}{P(Y=0|x=EX)} = \frac{P1}{1-P1}$, where $P1$ is defined as the probability that $Y = 1$ given $x =$the expected value of X

- We regress Y on X using the data of

$$\begin{array}{cc} X & Y \\ \left(\begin{array}{c} \\ \sim N(\mu_1, \sigma^2) \\ \\ \\ \sim N(\mu_2, \sigma^2) \\ \\ \end{array}\right. & \left.\begin{array}{c} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{array}\right) \end{array}$$

to fit the logistic model.

- $P(Y = 1|x) = \frac{f(x|Y = 1)P(Y=1)}{f(x)}$

- $\frac{P(Y = 1|x)}{P(Y = 0|x)} = \frac{f(x|Y = 1)P(Y=1)}{f(x|Y = 0)P(Y=0)} = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)P(Y=1)}{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)P(Y=0)}$

- $\log\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \frac{(x-\mu_1)^2-(x-\mu_2)^2}{2\sigma^2}$

$$= C + \frac{\mu_2 - \mu_1}{\sigma} \cdot \frac{x}{\sigma}$$

- So that we have

$$\frac{\mu_2-\mu_1}{\sigma} = \beta_1,$$

where β_1 has the meaning of the log OR when x increase to $x + \sigma$.

- Thus, $n = \left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \left[\frac{(k+1)^2}{k}\right] / (\frac{\mu_1-\mu_2}{\sigma})^2$

$$= \left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 / \left[P1(1-P1){\beta_1}^2\right]$$

$$= \left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 / \left[P1(1-P1)\log(\frac{\mathrm{OR}_2}{\mathrm{OR}_1})^2\right]$$

Parameters to be input in PASS

Data

Solve For

Find (Solve For): N

Error Rates

Power (1-Beta): 0.90

Alpha (Significance Level): 0.05

Sample Size

N (Sample Size): 20 to 300 by 50

Effect Size

Baseline Probability

P0 (Baseline Probability that Y=1): 0.3

Alternative Probability

Use P1 or Odds Ratio: P1

P1 (Alternative Probability that Y=1): 0.6

Odds Ratio (Odds1/Odds0): 1.5

Covariates (X1 is the Variable of Interest)

R-Squared of X1 with Other X's: 0

X1 (Independent Variable of Interest): Continuous (Normal)

Percent of N with X1=1: 50

Test

Logistic Regression Power Analysis

Numeric Results

Power	N	P0	P1	Odds Ratio	R Squared	Alpha	Beta
0.89185	31	0.300	0.600	3.500	0.000	0.05000	0.10815

References

Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. 'A Simple Method of Sample Size Calculation for Linear and Logistic Regression', Statistics in Medicine, Volume 17, pages 1623-1634.

Report Definitions

Power is the probability of rejecting a false null hypothesis. It should be close to one.
N is the size of the sample drawn from the population.
P0 is the response probability at the mean of X.
P1 is the response probability when X is increased to one standard deviation above the mean.
Odds Ratio is the odds ratio when P1 is on top. That is, it is [P1/(1-P1)]/[P0/(1-P0)].
R-Squared is the R2 achieved when X is regressed on the other independent variables in the regression.
Alpha is the probability of rejecting a true null hypothesis.
Beta is the probability of accepting a false null hypothesis.

Summary Statements

A logistic regression of a binary response variable (Y) on a continuous, normally distributed variable (X) with a sample size of 31 observations achieves 89% power at a 0.05000 significance level to detect a change in Prob(Y=1) from the value of 0.300 at the mean of X to 0.600 when X is increased to one standard deviation above the mean. This change corresponds to an odds ratio of 3.500.

When X is binary …

- The sample size formula for comparing the two independent event rates is:

$$n = (1+k) \cdot \left\{ Z_{1-\frac{\alpha}{2}} \left[\frac{p(1-p)(k+1)}{k} \right]^{\frac{1}{2}} + Z_{1-\beta} \left[p_0(1-p_0) + \frac{p_1(1-p_1)}{k} \right]^{\frac{1}{2}} \right\}^2 / (p_0 - p_1)^2$$

where p_0 and p_1 are proportion that Y=1 given X=0 or 1, and $p = \frac{p_0 + k p_1}{1+k}$

- If X is a binary variable, then $\beta_1 = 0$ if and only if the two event rates are equal.
- Suppose B is the proportion of the sample with $\mathrm{X} = 1$, p_1 is the event rate at $\mathrm{X} = 1$, p_0 is the event rate at $\mathrm{X} = 0$, overall event rate is $p = (1 - B)p_0 + Bp_1$. By replacing

$$k = \frac{\#\{X = 1\}}{\#\{X = 0\}} = \frac{P(X = 1)}{P(X = 0)} = \frac{B}{1 - B},$$

the formula becomes:

$$n = \left\{ Z_{1-\frac{\alpha}{2}} \left[\frac{p(1-p)}{B} \right]^{\frac{1}{2}} + Z_{1-\beta} \left[p_0(1-p_0) + \frac{p_1(1-p_1)(1-B)}{B} \right]^{\frac{1}{2}} \right\}^2 / [(p_0 - p_1)^2 (1 - B)]$$

where $p = (1 - B)p_0 + Bp_1$

For multiple logistic regression

- In multiple logistic regression, the hypothesis is:

$$H_0: [\beta_1, \beta_2, \dots, \beta_p] = [0, \beta_2, \dots, \beta_p]$$
$$H_1: [\beta_1, \beta_2, \dots, \beta_p] = [\beta_1^*, \beta_2, \dots, \beta_p]$$

- In the multivariate setting with p covariates, variance of β_1 's MLE b_1 $\mathrm{var_p(b_1)}$ can be approximated by inflating the variance of b_1 obtained from the one parameter model:

$$var_p(b_1) = var_1(b)/(1 - R^2)$$

where R^2 is equal to the proportion of the variance of $\mathrm{X_1}$ explained by the regression relationship with $\mathrm{X_2, \dots, X_p}$.

- The sample size can also be approximated from the univariate case by inflating it with the same factor:

$$n_p = n_1/(1 - R^2)$$