#9-2017F

Logistic Regression II

Summary of the last lecture

- Logistic regression: Terminology
- Ungrouped binary/Grouped binary

Key terms of this lecture

- Logistic regression (cont'd)
  - Other link function
  - Confounding and interaction
  - Different study designs

Reading

- McCullagh and Nelder (1989) Chapter 7
- Dobson and Bartnett (2008) Chapter 4

Other Link Functions

- Recall: logistic regression

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}_i^T\beta, \quad \text{for } i = 1, \quad , n.$$

  - Link function:

$$\eta = g(p) = \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

    i.e.

$$p = \frac{\exp(\eta)}{1+\exp(\eta)}$$

  - Very common way to model success probability p.

- It is related to the cdf of the logistic distribution:

$$\text{cdf} \qquad F(x) = \frac{\exp\{(x-\mu)/\tau\}}{1+\exp\{(x-\mu)/\tau\}}$$

$$\text{pdf} \qquad f(x = \frac{(1/\tau)\exp\{(x-\mu)/\tau\}}{[1+\exp\{(x-\mu)/\tau\}]^2}.$$

where $E[X] = \mu$ and $V[X] = \pi^2\tau^2/3$, and the pdf is symmetric and bell-shaped.

pdf is symmetric: e.g., Let $\mu=0$, $\tau=1$

$$f(x) = \frac{exp(x)}{\{1+exp(x)\}^2}$$

$$f(-x) = \frac{exp(-x)}{\{1+exp(-x)\}^2} = \frac{exp(-x)\ exp(2x)}{\{1+exp(-x)\}^2\ exp(2x)}$$

$$= \frac{exp(x)}{\{1+exp(x)\}^2}$$

$$= f(x)$$

Other Link (cont'd)

- Probit model
  - A natural alternative of logit distribution is normal distribution.
    - $p = \Phi(\eta)$, where Φ is a normal cdf
    - $\eta = g(p) = \Phi^{-1}(p) \rightarrow$ probit link function.
    - Program: `R: family=binomial(link=probit)`
      `SAS: model / dist=bin link=probit;`
  - Note:
    - Probit model was popularized by Bliss (1934,1935) for toxicological experiments prior to logistic model. Logistic model became more popular due to its interpretation of odds ratio (Agresti, Categorical Data Analysis, 1990)
    - MN Section 4.3.1 for a discussion of link functions.

In fact, for any continuous distribution function $F(x)$, let $p = F(\eta)$, then $\eta = F^{-1}(p)$, let $g(p) = F^{-1}(p)$ be the inverse function of F, then $g(p) = \eta$, $g(\cdot)$ is a link function in GLM.

Other Link (cont'd)

- Models with Complementary log-log link:
  - Let F be the cdf of the extreme value distribution (Gumbel distribution)
    - $p = F(\eta) = 1 - \exp\{-\exp(\eta)\}$.
    - $\eta = g(p) = \log\{-\log(1-p)\} \rightarrow$ Complementary log-log link.
    - Program: `R: family=binomial(link=cloglog)`
      `SAS: model / dist=bin link=cloglog;`
  - Note:
    - Gumbel distribution not symmetric
    - Interpretation of parameters is difficult
    - The model corresponds to the Cox's proportional hazards model for survival data. For any r.v. $T \sim$ Cox PH model, given x $\Lambda(t, x) = \Lambda_0(t)\exp(x^T\beta)$ $\Lambda_0(t)$ is the cumulative baseline hazard fun. Then $\log\Lambda(T, x) = \log\Lambda_0(T) + x^T\beta$, and $\log\Lambda_0(T)$ is a r.v. and $\sim$ Extreme Value distribution. In fact, if $x=0$, $S_0(T) = \exp(-\Lambda_0(T)) \sim U[0,1]$ (uniform dist) $\Lambda_0(T) = -\log(S_0(T)) \sim -\log U \sim \exp(\lambda=1)$ To show this, $\forall\ u \in (0, +\infty)$, $P(-\log U \le u) = P(U \ge e^{-u}) = 1 - e^{-u}$, Cdf of $Exp(\lambda=1)$.

STAT 635-GLM-Lecture Notes 9, Binary Variables and Logistic Regression, Part II, Fall 2017 We show

$\log\Lambda_0(T) = \log \exp(\lambda=1) \sim$ Extreme Value Dist

In fact, $\forall -\infty < \eta < \infty$,

$P = P(\log\Lambda_0(T) \le \eta)$
$= P(\Lambda_0(T) \le \exp(\eta))$
$= 1 - e^{-u}\big|_{u=\exp(\eta)}$
$= 1 - \exp(-\exp(\eta))$

Then $1 - p = \exp(-\exp(\eta))$
$\log(1-p) = -\exp(\eta)$
$-\log(1-p) = \exp(\eta)$
$\log\{-\log(1-p)\} = \eta$

Other Link (cont'd)

- log link:
  - Program: `R: family=binomial(link=log)`
    `SAS: model / dist=bin link=log;`
    Thus,
    $$\log(p_i) = \mathbf{X}_i^T\beta.$$
  - Relative risk interpretation
    - Consider the following model
      $$\log(p_i) = \beta_0 + \beta_1 X_i.$$
    - The relative risk at $x+1$ vs. x given by
      $$RR = \frac{p_{x+1}}{p_x} = \frac{P(Y=1|X=x+1)}{P(Y=1|X=x)} = \frac{\exp(\beta_0+\beta_1(x+1))}{\exp(\beta_0+\beta_1 x)} = \exp(\beta_1)$$
      $\longrightarrow$ Comp. log-log link.
  - No guarantee that p_i will range between 0 and 1.

To see this, notice $p = \exp(x^T\beta) \in (0, +\infty)$, Since $x^T\beta \in (-\infty, \infty)$

Relative Risk and Odds Ratio

- For an unmatched case-control study, the data look like this:

Exposed	Cases (D)	Controls ($\bar{D}$)	Total
Yes	a	b.	a+b
No	c	d.	c+d
Total	a+c	b+d	a+b+c+d

- **RR** The relative risk (RR) is the probability that a member of an exposed (E) group will develop a disease relative to the probability that a member of an unexposed ($\bar{E}$) group will develop that same disease. It can be estimated from a ~~retrospective~~ prospective cohort study where $a+b$ and $c+d$ are fixed.

$$RR = \frac{P(D|E)}{P(D|\bar{E})} = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}$$

- **Odds ratio: OR** Odds ratio of disease:

$$OR_D = \frac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))} = \frac{\frac{a}{a+b} / \frac{b}{a+b}}{\frac{c}{c+d} / \frac{d}{c+d}} = \frac{ad}{bc}$$

  Used in a prospective cohort study.

- **Odds ratio: OR** Odds ratio of exposure:

$$OR_E = \frac{P(E|D)/(1-P(E|D))}{P(E|\bar{D})/(1-P(E|\bar{D}))} = \frac{\frac{a}{a+c} / \frac{c}{a+c}}{\frac{b}{b+d} / \frac{d}{b+d}} = \frac{ad}{bc}$$

  Used in a a case-control study, where the number of cases $a+c$ and controls $b+d$ are fixed.

- Note that these two definitions are equivalent, i.e., $OR_D = OR_E$.
- For rare disease, $P(D|\bar{E})$ and $P(D|\bar{E})$ are small, so $RR \approx OR$, OR can be used to estimate RR from a case-control study.

Confounding

- When some association between a covariate and the outcome is observed, we often ask two questions:
  1. Is the association real?
     * It might be false because of all sorts of **biases**.
     * Selection bias: diseased and non-diseased subjects are sampled on the basis of differing criteria, and these criteria are related to the exposure. (e.g.,case-control study)
     * Recall bias: diseased and non-diseased subjects may have differential recall about exposures.
     * Loss-to-follow-up: diseased and non-diseased subjects are lost via an underlying mechanism that is related to exposure.
     * Misclassification bias.

  2. If the observed association is real, then is it causal (i.e. cause-and-effect)?
     * An observed association may be due to confounding, one of the most important problems in observational studies. This is different from the designed experiments, where confounding can be controlled

Confounding (cont'd)

- **Confounding.**
  - In a study of whether factor A is a cause of disease B, we say that a third factor X is a confounder if the followings are true:
    - X is a risk factor of disease B.
    - X is associated with A but not a result of A.

A → B, X (confounder) — Association

Definition: A lurking variable is a variable that is not among the explanatory or response variable in a study and yet may influence the interpretative of relationship among theses variables.

  - Example (Hypothetical)

Exposed	Cases	Controls
Yes	30	18
No	70	82
Total	100	100

Definition of Confounding: Two variables are confounded when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variable or lurking variable.

* OR of Cases:

$$\hat{OR} = \frac{ad}{bc} = \frac{30 \times 82}{70 \times 18} = \frac{2460}{1260} = 1.9524$$

* How to establish a direct link between A and B? The best method is to conduct a carefully designed experiment in which the effects of possible lurking variables are controlled.
* Association vs Causal effect. Strong association does not imply a causal effect

STAT 635-GLM-Lecture Notes 9, Binary Variables and Logistic Regression, Part II, Fall 2017

Confounding (cont'd)

OR (of cases at (age < 40) relative to (age ≥ 40)) $= \frac{50 \times 20}{50 \times 80} = 0.25$

  - Example (cont'd): There is another factor Age.

Exposure is a risk factor (A), it may be a cause of disease (B). Age is also a risk factor (X). Age is associated with Exposure but not a result of Exposure. See two tables below.

Age	Cases	Controls
< 40	50	80
≥ 40	50	20
Total	100	100

Age is a Confounder

* Age < 40:

Exposed	Cases	Controls
Yes	5	8
No	45	72
Total	50	80

What is OR of Cases? $OR_{<40} = \frac{5 \times 72}{45 \times 8} = \frac{360}{360} = 1$

* Age ≥ 40:

Exposed	Cases	Controls
Yes	25	10
No	25	10
Total	50	20

What is OR of Cases? $OR_{\geq 40} = \frac{25 \times 10}{25 \times 10} = 1$

Both are different from $OR = 1.9524$ in the pooled data analysis

Confounding (cont'd)

- How to handle confounding?
  - In designing and carrying out the study:
    * Matching.
    * e.g., case-control study: age is a probable confounding, then each 60 years old patient (case) matched with a healthy 60 year old person (control).
  - In the statistical analysis:
    * Stratification.
    * Adjustment: find potential confounders and include them in the model.

Study Designs

- **Intervention**: Investigator controls assignment of exposure prospectively.
  - Gold standard of study design since it allows to control confounders.
  - Issue: failure to comply, ethical requirements..
- **Cohort**: Sample of *exposed* and *unexposed* who are both at risk for the disease are followed for a period of time to determine disease status; can be prospective or retrospective.
  - Interested in incidence rate (i.e. counts of disease outcomes over person-years of exposure) $\rightarrow$ Poisson regression.
  - Need to adjust for potential confounders.
  - Loss-to-follow up bias.

- **Case-Control**: Sample of *diseased* and *nondiseased* who could have had the exposure of interest have exposure status ascertained retrospectively.
  - Interested in OR
  - Selection bias; potential confounders