BIOSTAT 651

Notes #10: Case-Control & Link functions

- Lecture Topics:
  - Case-control sampling
  - Link functions

Case-Control sampling

Study Designs

- Study designs:
  - randomized clinical trial
  - observational study
- Randomized trial: gold standard
  - often infeasible (logistics, ethics)
- Observational study: often easier and more cost efficient to study associations
  - cohort
  - case-control

Case-Control Sampling

- Case-Control study:
  - select n_1 diseased subjects
  - select n_0 non-diseased subjects
  - analysis: contrast $\mathbf{x}_i$ between *cases* ($Y_i = 1$) and *controls* ($Y_i = 0$)
- Motivation: study of *rare* diseases
  - more generally, *cost*- and/or *time-efficient* study of disease (more later...)

Analysis of Case-Control Data

- Recall: the exposure odds ratio (EOR) estimated through case-control sampling is equal to the OR of interest
  - i.e., $Y_i = 0, 1$ and $X_i = 0, 1$

$$
\begin{aligned}
EOR &\equiv \frac{\text{odds}(X_i = 1 | Y_i = 1)}{\text{odds}(X_i = 1 | Y_i = 0)} \\
&= \vdots \\
&= \vdots \\
&= \frac{\text{odds}(Y_i = 1 | X_i = 1)}{\text{odds}(Y_i = 1 | X_i = 0)} \equiv OR
\end{aligned}
$$

- Result extends to the regression setting...

Case-Control Study: General Set-Up

- Consider the following setting:
  - Y_i: binary
  - $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{iq})^T$ (assume discrete)
  - population: N subjects; study: n subjects

    sample n_1 cases, and n_0 controls

    S_i = sampling indicator
- data:
  - observed (population):
  - assigned (by investigators):
  - observed data (sample):

- Model:

∘ sampling fractions:

$$\tau_0 \equiv P(S_i = 1 | Y_i = 0)$$

$$\tau_1 \equiv P(S_i = 1 | Y_i = 1)$$

∘ model:

$$\begin{aligned} \log\left\{\frac{\pi_i}{1-\pi_i}\right\} &= \mathbf{x}_i^T \boldsymbol{\beta} \\ \pi_i = \pi(\mathbf{x}_i) &= \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \end{aligned}$$

Case-Control Data: MLE

- Estimation proceeds via maximum likelihood
  - Since $\mathbf{x}_i$ is a random variable, the retrospective likelihood is written as

$$L_i(\boldsymbol{\beta}) \propto P(\mathbf{x}_i|Y_i = 1, S_i = 1)^{Y_i}$$
$$P(\mathbf{x}_i|Y_i = 0, S_i = 1)^{1-Y_i}$$

  - Prentice and Pyke (1979) showed that MLE of β from $P(Y_i = 1|\mathbf{x}_i, S_i = 1)$ is the same as MLE of β from the prospective logistic regression model with retrospective sampling.

Case-Control Data: MLE

- $P(Y_i = 1|\mathbf{x}_i, S_i = 1)$:

$$\begin{aligned} & P(Y_i = 1|\mathbf{x}_i, S_i = 1) \\ = \quad & \frac{P(Y_i = 1, \mathbf{x}_i, S_i = 1)}{P(\mathbf{x}_i, S_i = 1)} \\ = \quad & \frac{P(\mathbf{x}_i|Y_i = 1, S_i = 1)P(Y_i, S_i = 1)}{P(\mathbf{x}_i, S_i = 1)} \qquad (1) \end{aligned}$$

- Since

$$P(S_i = 1|\mathbf{x}_i, Y_i = 1) = P(S_i = 1|Y_i = 1),$$

we have

$$\begin{aligned} & P(\mathbf{x}_i|Y_i = 1, S_i = 1) \\ = \quad & \frac{P(S_i = 1|\mathbf{x}_i, Y_i = 1)P(\mathbf{x}_i|Y_i = 1)}{P(S_i = 1|Y_i = 1)} \\ = \quad & P(\mathbf{x}_i|Y_i = 1) \\ = \quad & \frac{P(Y_i = 1|\mathbf{x}_i)P(\mathbf{x}_i)}{P(Y_i = 1)} \qquad (2) \end{aligned}$$

- Combine (1) and (2)

$$
\begin{aligned}
P(Y_i = 1|\mathbf{x}_i, S_i = 1) &= P(Y_i = 1|\mathbf{x}_i)\frac{P(S_i = 1|Y_i = 1)}{P(S_i = 1|\mathbf{x}_i)} \\
&= \pi(\mathbf{x}_i)\frac{P(S_i = 1|Y_i = 1)}{P(S_i = 1|\mathbf{x}_i)}
\end{aligned}
$$

- Odds

$$
\begin{aligned}
\frac{P(Y_i = 1|\mathbf{x}_i, S_i = 1)}{P(Y_i = 0|\mathbf{x}_i, S_i = 1)} &= \frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\frac{P(S_i = 1|Y_i = 1)}{P(S_i = 1|Y_i = 0)} \\
&= \frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\frac{\tau_1}{\tau_0}
\end{aligned}
$$

- Odds ratio between $\mathbf{x}_a$ vs $\mathbf{x}_b$

$$
OR = \frac{\pi(\mathbf{x}_a)/\{1-\pi(\mathbf{x}_a)\}}{\pi(\mathbf{x}_b)/\{1-\pi(\mathbf{x}_b)\}}
$$

OR is the same as the prospective design OR.

- Logistic regression model with retrospective sampling

$$\begin{aligned}\log\{odds(\mathbf{x}_i)\} &= \mathbf{x}_i'\beta^* \\ &= \mathbf{x}_i'\beta + \log\left(\frac{\tau_1}{\tau_0}\right)\end{aligned}$$

- Intercept:

$$\beta_0^* = \beta_0 + \log\left(\frac{\tau_1}{\tau_0}\right)$$

  - Only the intercept is changed.
  - If the sampling fractions are known (τ_0 and τ_1), we can estimate true β_0.

Case-Control: Example

- We return to the lung cancer example in Lecture Note #8 . . .
  - Smoking is a risk factor for the colorectal cancer. It can increase the risk twice.
  - Assumptions:
    - Risk for the cancer among non-smoker: 0.05
    - Prevalence of smoking: 20%
  - Studies
    - Case-control study with 5,000 cases vs 5,000 controls.

- Consider a saturated model based on the logit link,

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} \quad = \quad \beta_0 + \beta_1 X_i$$

  - True values

$$\begin{aligned}
\beta_0 &= \log\left\{\frac{0.05}{0.95}\right\} = -2.944 \\
\beta_0 + \beta_1 &= \log\left\{\frac{0.1}{0.9}\right\} = -2.197 \\
\beta_1 &= 0.747; \quad \exp(\beta_1) = 2.11
\end{aligned}$$

Example : Case-Control Sampling (logit)

	Y=0	Y=1	total
X=0	4022	3358	7380
X=1	978	1642	2620
total	5000	5000	10000

- we compute the parameter estimates as:

$$
\begin{aligned}
\widehat{\beta}_0 &= \log\left\{\frac{3358}{4022}\right\} = -0.1804 \\
\widehat{\beta}_0 + \widehat{\beta}_1 &= \log\left\{\frac{1642}{978}\right\} = 0.518 \\
\widehat{\beta}_1 &= 0.698; \quad \exp(\widehat{\beta}_1) = 2.01
\end{aligned}
$$

- Calculate sampling fraction:
  - Prevalence:

$$P(Y = 1) = 0.06$$

  - Ratio of the sampling fractions:

$$\frac{\tau_1}{\tau_0} = 0.94/0.06 = 15.6667$$

- Adjust β_0 using the sampling fractions:

$$\widehat{\beta}_0 - \log(\frac{\tau_1}{\tau_0}) = -0.1804 - 2.7515 = -2.9319$$

Outcome-Dependent Sampling

- Case-control study is a special case of what has come to be called *Outcome-Dependent Sampling* (ODS)
    - creative ways to sample cases and controls
    - outcome can be binary, or more complicated structure
    - extension to survival times, clustered data, etc

Link functions

- Dose-response modeling
- Link functions
- Interpretation of parameters
- Issues in case-control sampling

Dose-Response Models: Introduction

- General set-up:
  - evaluate effect of dose on the probability of a specific event
  - covariate: D_i
  - dependent variable: $Y_i = 0, 1$
  - set $\pi_i = \pi(D_i) = P(Y_i = 1|D_i)$
- Based on our study to date, we use the logit link
  - we now explore alternative link functions

Link Functions: Binary Response

- Link functions used for binary data:

$$
\begin{aligned}
\log\left\{\frac{\pi_i}{1-\pi_i}\right\} &= \mathbf{x}_i^T\boldsymbol{\beta} \\
\Phi^{-1}(\pi_i) &= \mathbf{x}_i^T\boldsymbol{\beta} \\
\log\{-\log(1-\pi_i)\} &= \mathbf{x}_i^T\boldsymbol{\beta} \\
-\log\{-\log(\pi_i)\} &= \mathbf{x}_i^T\boldsymbol{\beta}
\end{aligned}
$$

  - all are continuous and increasing on $(0, 1)$
  - only first two are symmetric

Link Functions: Background

- Often model π_i using a cumulative distribution function

$$\pi(t) \quad = \quad \int_{-\infty}^{t} f(s)ds$$

where $f(s)$ is a *tolerance* distribution

- ◦ characteristics of valid $f(s)$:

(i) $f(s) \geq 0$

(ii) $\int_{-\infty}^{\infty} f(s)ds = 1$

Connection to Bioassays

- Historically, binomial regression models were motivated by *bioassay* studies
  - response: proportion of events

    e.g., percent dead
  - exposure: dose level

    e.g., treatment, toxin, contaminant, etc
- Models of the form $g(\pi_i) = \beta_0 + \beta_1 D_i$ were often considered

Example: Uniform Tolerance

- Example: Suppose that the tolerance distribution is Uniform(a, b):

$$f(s) =$$

$$\pi(t) =$$

  - Graphs of $f(s)$ and $\pi(t)$:

Uniform Tolerance

- Connecting dose and tolerance:

$$
\begin{aligned}
\pi(x) &= \beta_0 + \beta_1 x \\
\beta_0 &= \frac{-a}{b-a} \\
\beta_1 &= \frac{1}{b-a}
\end{aligned}
$$

- Need to impose constraints on x, β_0 and β_1
  - standard GLM methods would not apply

Probit Model

- Suppose that a Normal(μ, σ^2) distribution is used for the tolerance
  - $f(s) =$
  - $P(T \leq t) =$

- Choosing the probit function as the link:

$$
\begin{aligned}
\Phi^{-1}(\pi(x)) &= \beta_0 + \beta_1 x \\
\beta_0 &= -\frac{\mu}{\sigma} \\
\beta_1 &= \frac{1}{\sigma}
\end{aligned}
$$

Probit Model (continued)

- Probit link is used frequently
  - e.g., $Y_i = I_i(\text{dead})$

    μ referred to as the median lethal dose,

    LD(50): dose required to kill 50% of the members of a tested population in a specific time.

Logit Link

- Consider the logistic tolerance distribution:

$$f(s) \quad = \quad \frac{\beta_1 \exp\{\beta_0 + \beta_1 s\}}{(1 + \exp\{\beta_0 + \beta_1 s\})^2}$$

which implies that

$\pi(t) =$

which implies the *logit* link function

Complementary Log-Log Link

- If tolerance follows the *extreme value* distribution:

$$f(s) \quad = \quad \beta_1 \exp\{(\beta_0 + \beta_1 s) - e^{\beta_0 + \beta_1 s}\}$$

then we obtain

$$\pi(t) \quad = \quad 1 - \exp\left\{-e^{\beta_0 + \beta_1 t}\right\}$$

which implies the *complementary log-log* link:

$$\log\{-\log(1 - \pi(x))\} \quad = \quad \beta_0 + \beta_1 x$$

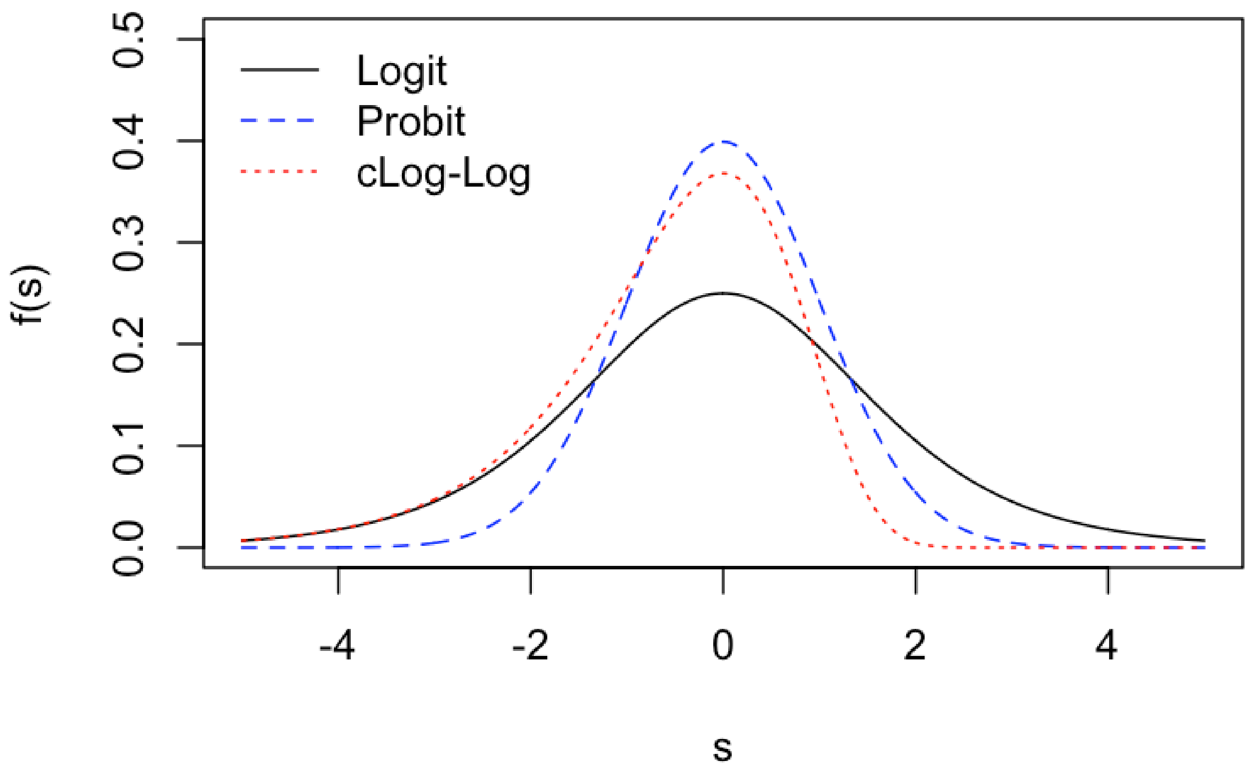
- Related to the harzard ratio
  - Hazard function

$$h(t) = P(T = t | T \geq t) = \frac{f(t)}{1 - \pi(t)}$$

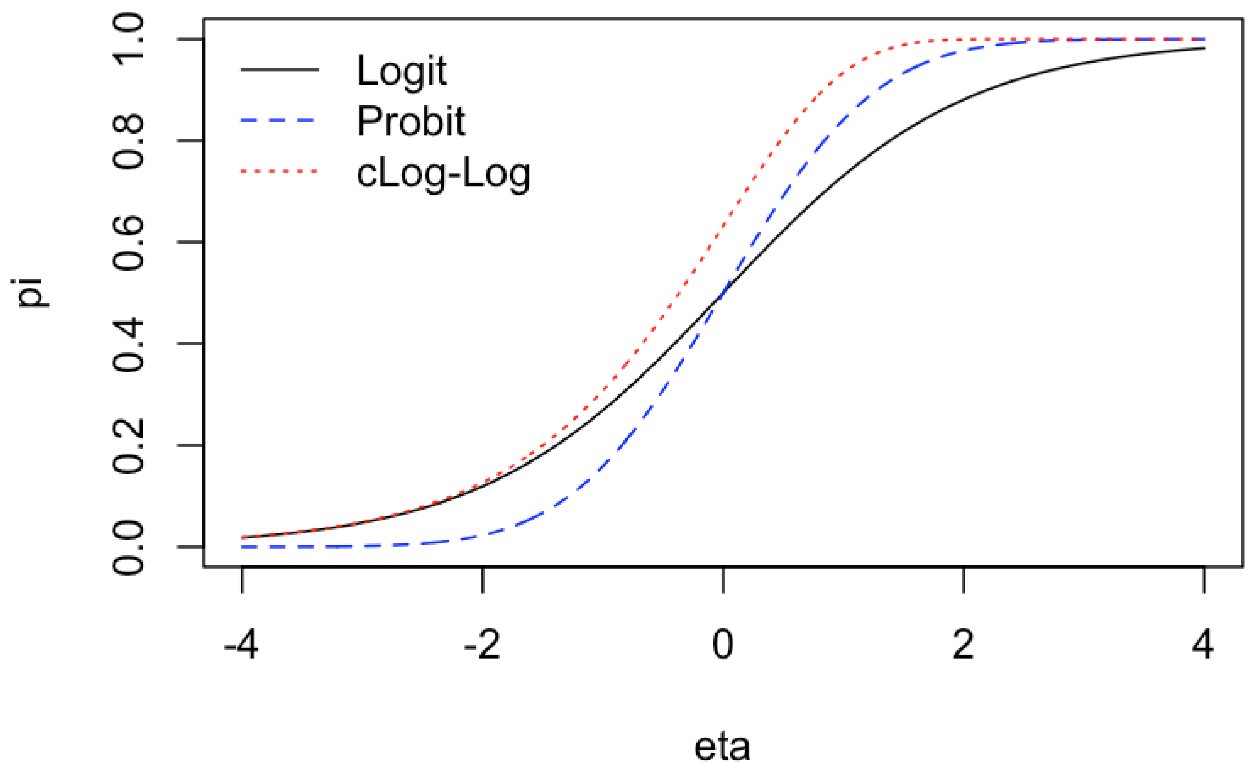
  - Hazard ratio:

$$\frac{h(t+1)}{h(t)} = \exp(\beta_1)$$

$f(s)$ functions with $\beta_0 = 0$ and $\beta_1 = 1$



$\pi(x)$ functions



Case-Control Sampling

- Recall that logistic regression provides a consistent OR estimator for case-control sampling
  - e.g, if the model is given by

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_0 + \mathbf{x}_{i1}^T\boldsymbol{\beta}_1$$

    then a case-control study will consistently estimate:

- This is a property of the logit link
  - need not hold for alternative link functions

Case-Control Sampling: Example

- Lung cancer example
  - Cohort study with 5,000 non-smoker vs 5,000 smokers
  - Case-control study with 5,000 cases vs 5,000 controls.

Example: Complementary log-log link

- Example: Assume that the true model follows the complementary log-log link,

$$\log\{-\log(1-\pi_i)\} = \beta_0 + \beta_1 X_i$$

- True values

$$\beta_0 = \log\{-\log(1-0.05)\} = -2.97$$

$$\beta_0 + \beta_1 = \log\{-\log(1-0.1)\} = -2.25$$

$$\beta_1 = 0.720$$

Example: Cohort Sampling (CLL)

	$Y=0$	$Y=1$	total
$X=0$	4748	252	5000
$X=1$	4465	535	5000
total	9213	787	10000

- Parameter estimates:

$$\begin{aligned}
\widehat{\beta}_0 &= \log\{-\log(4748/5000)\} = -2.962 \\
\widehat{\beta}_0 + \widehat{\beta}_1 &= \log\{-\log(4465/5000)\} = -2.178 \\
\widehat{\beta}_1 &= 0.783
\end{aligned}$$

Example : Case-Control Sampling (CLL)

	$Y=0$	$Y=1$	total
$X=0$	4022	3358	7380
$X=1$	978	1642	2620
total	5000	5000	10000

- Parameter estimates:

$$
\begin{aligned}
\widehat{\beta}_0 &= \log\{-\log(4022/7380)\} = -0.499 \\
\widehat{\beta}_0 + \widehat{\beta}_1 &= \log\{-\log(978/2620)\} = -0.0147 \\
\widehat{\beta}_1 &= 0.484
\end{aligned}
$$