BIOS 7345: Advanced Regression Analysis I

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 18: Further considerations for binary outcomes

Version: 09/17/2023

Table of Contents

Non-collapsibility of the logit link

Logistic regression: Non-collapsibility

- Take, for example, a study to determine the association between cardiovascular disease (CVD) and kidney stone history (yes/no).
- Tables stratified by age group ($\leq$50 years and $>$50 years):

$\leq$50 years	CVD	No CVD
Kidney stones	40	60
No kidney stones	10	90

$>$50 years	CVD	No CVD
Kidney stones	90	10
No kidney stones	60	40

- Is age group a confounder?
- What is the within-group (adjusted) odds ratio?

Non-collapsibility of the logit link

Logistic regression: Non-collapsibility

- Now, "collapse" the tables over age group:

	All	CVD	No CVD
Kidney stones	130		70
No kidney stones	70		130

- What is the pooled/unadjusted/crude odds ratio?

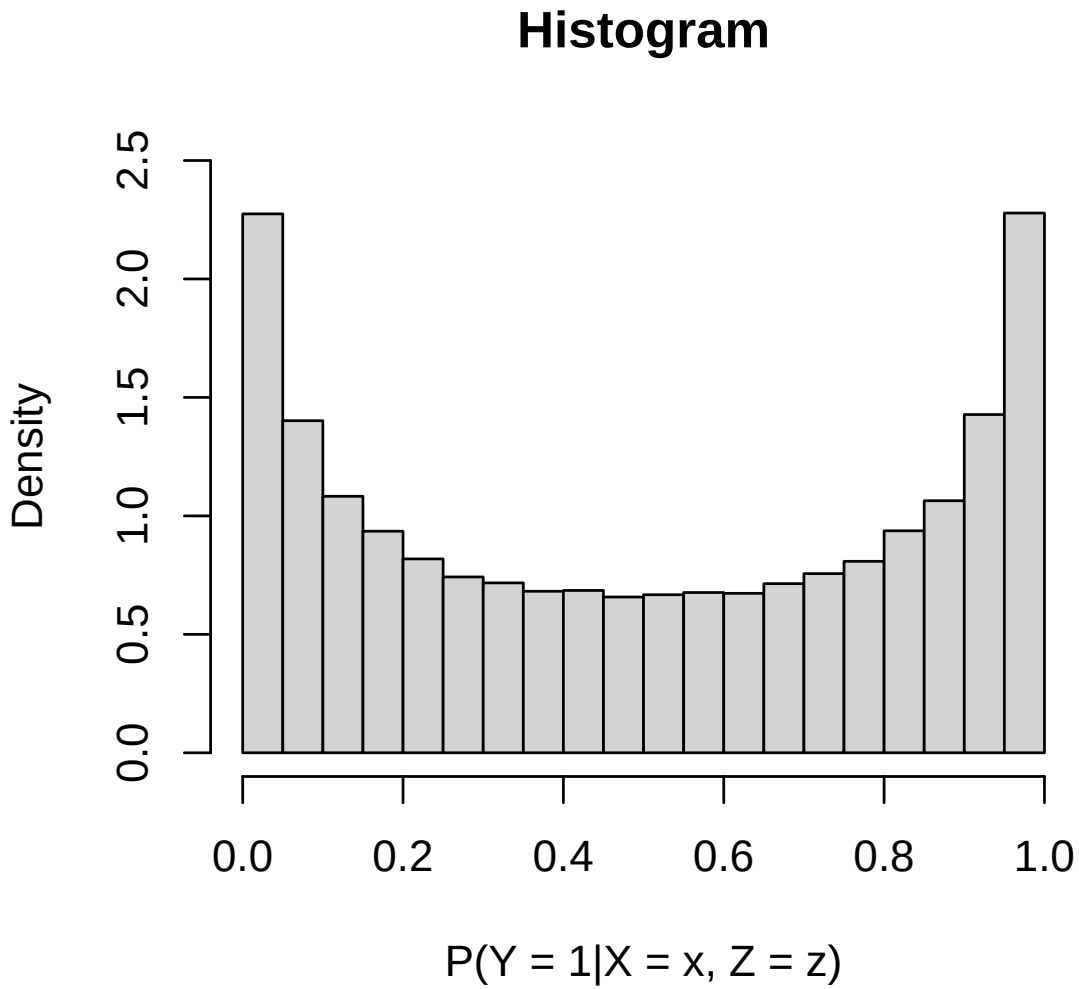
Non-collapsibility of the logit link

Illustration:

```
expit <- function(x)
{
  expitx <- exp(x)/(1 + exp(x))
  return(expitx)
}

## Very large sample size (answer close to truth)
n <- 100000

## Set seed
set.seed(7345)

## Generate data
X <- rnorm(n, 0, 1)
Z <- rnorm(n, 0, 3)
p <- expit(0.4*X + 0.8*Z)
Y <- rbinom(n, 1, p)
```

Non-collapsibility of the logit link

Logistic regression: Heterogeneity of $P(Y = 1|X = x, Z = z)$



Non-collapsibility of the logit link

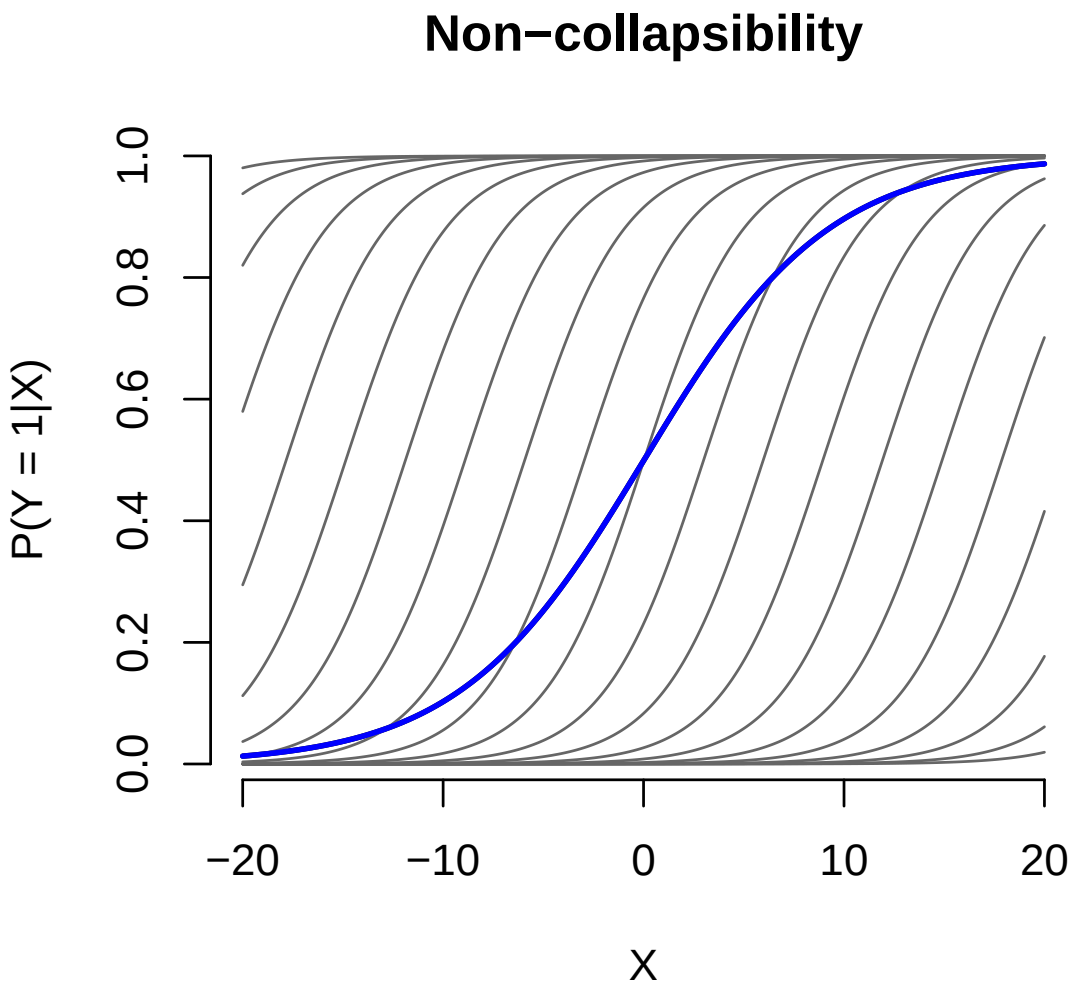
Illustration:

```
zz.1 <- glm(Y ~ X, family = binomial(link = "logit"))
zz.2 <- glm(Y ~ X + Z, family = binomial(link = "logit"))

b.1 <- as.numeric(coef(zz.1))
b.2 <- as.numeric(coef(zz.2))
```

```
> b.1
[1] -0.004947256  0.216420075
> b.2
[1] -0.009635699  0.401667754  0.796640094
```

Non-collapsibility of the logit link

Logistic regression: Illustration of non-collapsibility



Non-collapsibility of the logit link

Logistic regression: Non-collapsibility

- This illustration was a little "unfair" because only the model conditional on Z was actually correctly specified.
- The "fair" approach would be to consider a likelihood based on X that specifically marginalizes over the values of Z (beyond the scope of this course).
- Nevertheless, the estimate of a marginalized likelihood would be consistent for the conditional association (given Z).

Table of Contents

Latent variable formulation logistic regression

Logistic regression: Our typical formulations

- $Y \sim \text{Bernoulli}(p = \text{expit}(\mathbf{x}^T\boldsymbol{\beta}))$.
- $\text{logit}(\text{P}(Y = 1|\mathbf{X} = \mathbf{x})) = \mathbf{x}^T\boldsymbol{\beta}$.

Latent variable formulation logistic regression

Logistic regression: Latent variable formulation

- Imagine that for each Bernoulli trial, there is an underlying but latent continuous variable Y^* distributed as follows:

$$Y^* = \mathbf{x}^T \boldsymbol{\beta} + \epsilon,$$

where $\epsilon \sim \text{Logistic}(0, 1)$, with density:

$$f(\epsilon) = \frac{\exp(-\epsilon)}{(1 + \exp(-\epsilon))^2} = \frac{1}{4}\text{sech}^2(\epsilon/2).$$

- $Y = 1(Y^* > 0)$ marks the observed dichotomous outcome.
- Note: $\text{E}[\epsilon] = 0$ and $\text{Var}[\epsilon] = \pi^2/3$.

Latent variable formulation logistic regression

Logistic regression: Latent variable formulation

- The logistic family is a location-scale family, although there is no reason to allow ϵ to follow a more general form depending upon location and scale parameters:
  - Location is counteracted by the centering of the intercept, β_0.
  - Scale is counteracted by the magnitude of each component of $\boldsymbol{\beta}$.
- Latent variable formulation makes it easier to extend to more complicated models with multiple correlated outcomes as well as to compare logistic regression to the closely related probit model.

Table of Contents

The probit link function for binary outcomes

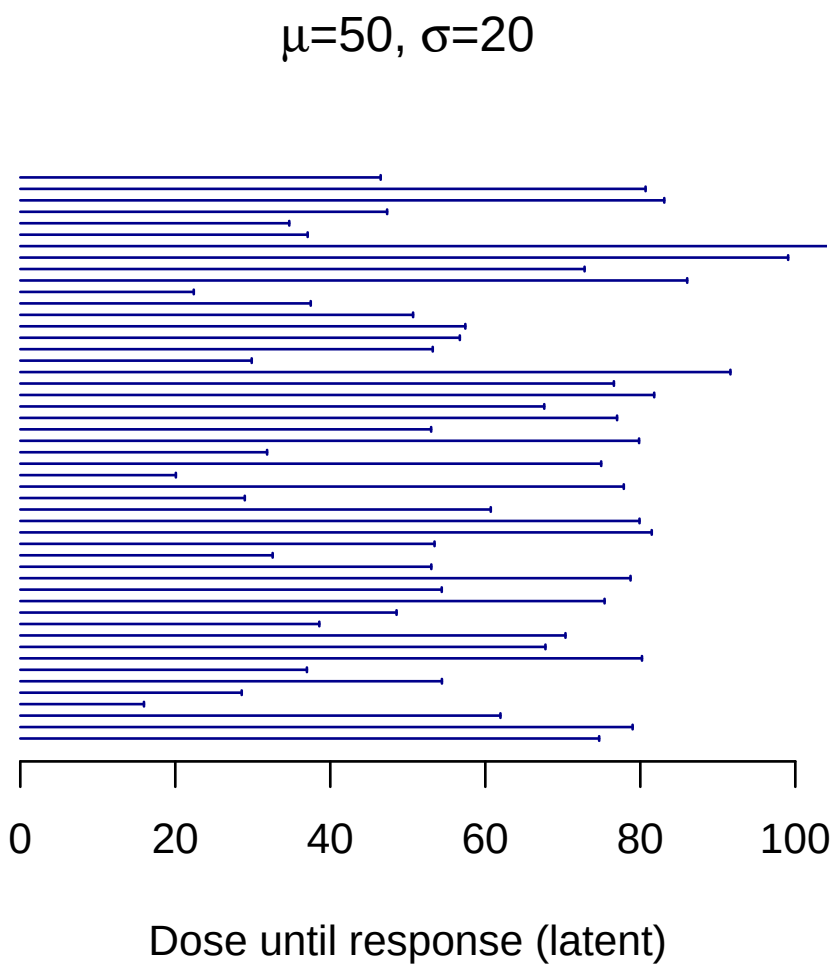
Example: Dose until response

- Consider a study in which you are seeking to determine the relationship between dose and probability of response.
- Imagine each study subject has an underlying dose until response:

$$Y^* \sim \mathcal{N}(\mu, \sigma^2).$$

The probit link function for binary outcomes

Example: Dose until response



The probit link function for binary outcomes

Example: Dose until response

- Each study unit receives one dose, $D = d$, and either responds ($Y = 1$) or not ($Y = 0$).
- Under the normality assumption, it follows that

$$\begin{aligned} \mathrm{P}(Y = 1|D = d) &= \mathrm{P}(Y^* < D|D = d) \\ &= \Phi\left(\frac{d - \mu}{\sigma}\right) \\ &= \Phi\left(-\frac{\mu}{\sigma} + \frac{1}{\sigma}d\right) \\ &= \Phi(\beta_0 + \beta_1 d). \end{aligned}$$

- By this model, μ and σ are not identifiable, but β_0 and β_1 are.
  - Groups differing by one unit in their received dose differ in their dose-to-response by β_1 standard deviations.

The probit link function for binary outcomes

Example: Dose until response

```
## Generate binary outcome accordingly
Y <- rbinom(n, 1, pnorm(D, mean = 50, sd = 25))

## Fit probit GLM
zz <- glm(Y ~ D, family = binomial(link = "probit"))

## Salient part of output
> summary(zz)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.96737    0.63920  -3.078 0.002085 **
D            0.04113    0.01099   3.741 0.000183 ***

## The true values are b0 = -50/25 = -2 and b1 = 1/25 = 0.04.
```

The probit link function for binary outcomes

Comparison: logit vs. probit

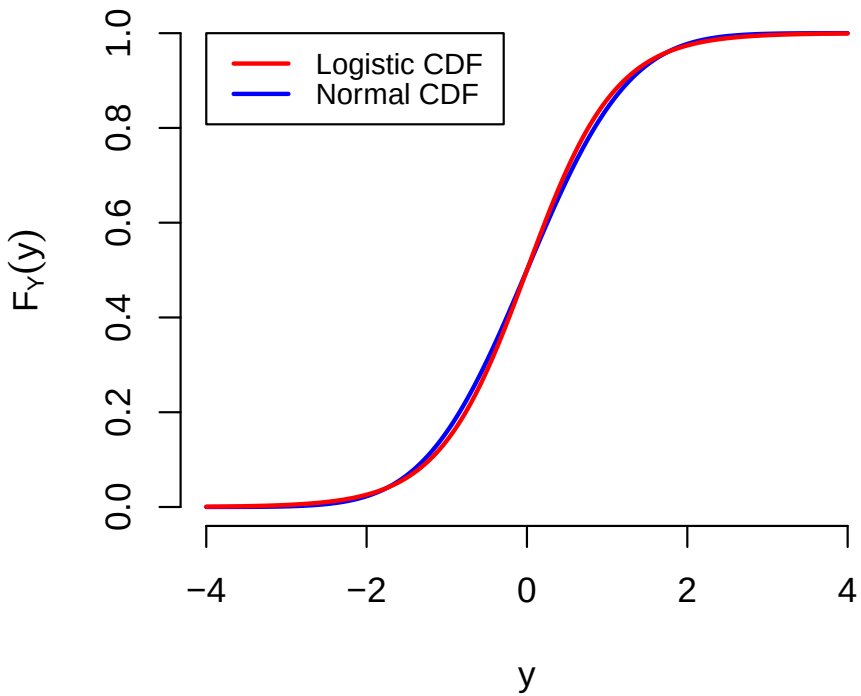


Table of Contents

The clog-log link function for binary outcomes

Another link function:

- Suppose we're interested in drawing inferences about the rates at which a binary outcome occurs.
- Suppose that Y^* denotes the time-to-event during some fixed period of follow-up, τ, that an event occurred (if it did):

$$\lambda(y^*) = \lim_{\delta \longrightarrow 0} \frac{\mathrm{P}(y^* \leq Y^* < y^* + \delta | Y^* \geq y^*)}{\delta}.$$

- If we assume $\lambda(y^*|X = x) = \lambda_0(y^*)\exp(\beta_1 x)$, then β_1 marks the hazard ratio that compares the instantaneous hazard rate of the event between subgroups differing in their value of X by one unit.

The clog-log link function for binary outcomes

Deriving the clog-log link:

- Let $Y = 1(Y^* < \tau)$ denote the observed indicator of having observed an event during the fixed period of follow-up.
- Noting that $f(y^*|X = x) = \lambda(y^*|X = x) \exp\left(-\Lambda(y^*|X = x)\right)$, we have:

$$1 - \mathrm{P}(Y = 1|X = 1) \overset{\text{math}}{=\!=\!=} (1 - \mathrm{P}(Y = 1|X = 0))^{\exp(\beta_1)}.$$

- That is to say that the proportional hazards assumption lends itself to the link function $g(\mu) = \log(-\log(1 - \mu))$.

The clog-log link function for binary outcomes

The clog-log link: $g(\mu) = \log(-\log(1-\mu))$

- Bernoulli GLM with the clog-log link:

$$\begin{aligned}
\log(-\log(1-\mathrm{P}(Y=1|X=x))) &= \beta_0+\beta_1 x \\
\Rightarrow 1-\mathrm{P}(Y=1|X=x) &= \exp(-\exp(\beta_0+\beta_1 x)) \\
&= \exp(-\exp(\beta_0))^{\exp(\beta_1 x)} \\
&= (1-\mathrm{P}(Y=1|X=0))^{\exp(\beta_1 x)}.
\end{aligned}$$

- Note: $\mathrm{P}(Y=1|X=0) = \mathrm{P}(Y^*<\tau|X=0) = 1-\exp(-\exp(\beta_0))$.
  - Note: This CDF is related to a standardized Gumbel distribution (latent variable formulation).
- Note: $\exp(\beta_1)$ is the hazard ratio.

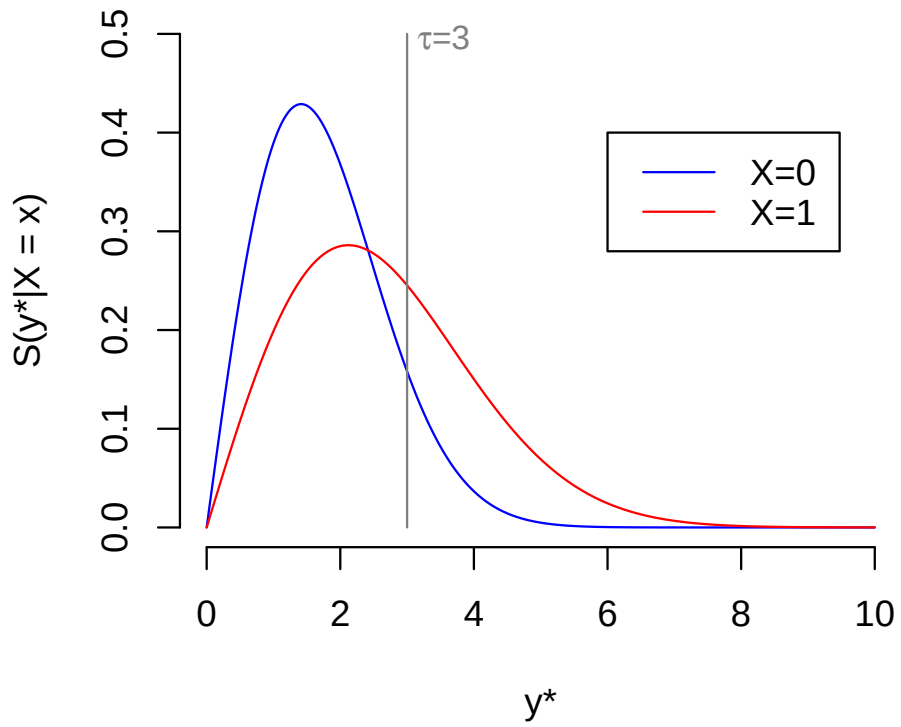
The clog-log link function for binary outcomes

Example: Weibull time-to-event with common shape

- Suppose $X \sim \text{Bernoulli}(p = 0.5)$ and $Y \sim \text{Weibull}(k = 2, \lambda = 2 + X)$.
- Consider the cutoff of $\tau = 3$.
  - $P(Y = 1|X = 0) \approx 0.895$ (`pweibull(3, shape = 2, scale = 2)`).
  - You can verify that this is consistent with the proportional hazards assumption, with a hazard ratio given by $(2/3)^2 = 4/9 \approx 0.444$, which should be true in general (and not only for this value of τ).

The clog-log link function for binary outcomes

Example: Weibull time-to-event with common shape



Example: Weibull time-to-event with common shape

```
## Set seed for reproducibility
set.seed(7345)

## Set sample size
n <- 2000

## Generate X
x <- rbinom(n, 1, 0.5)

## Generate underlying latent continuous variable
ystar <- rweibull(n, shape = 2, scale = 2 + x)

## Generate dichotomous realization with tau = 3
y <- as.numeric(ystar < 3)
```

The clog-log link function for binary outcomes

Example: Weibull time-to-event with common shape

```
## Fit clog-log GLM
zz <- glm(y ~ x, family = binomial(link = "cloglog"))

## Should be about 0.444
> as.numeric(exp(coef(zz)[2]))
[1] 0.456471

## Should be about 0.895
> as.numeric(1 - exp(-exp(coef(zz)[1])))
[1] 0.8849647
```

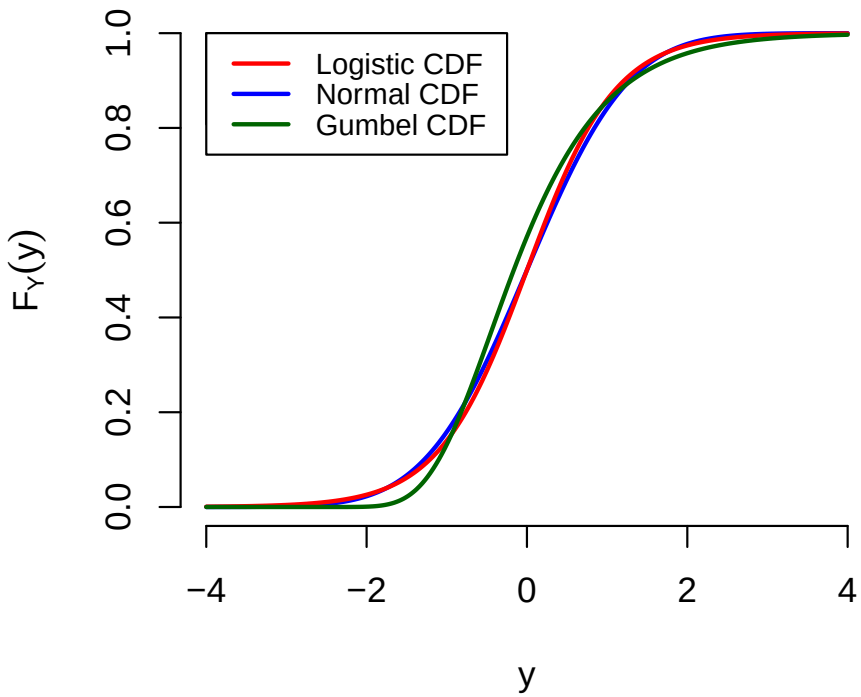
The clog-log link function for binary outcomes

Another link function:

- In this set of notes, we have characterized the logit, probit, and clog-log links for binary outcomes in terms of underlying/latent continuous variables.
- Generally, the choice of a link function for binary outcomes is more heavily driven by the interpretation you seek to get out of the model (and less by trying to fit the data perfectly).

The clog-log link function for binary outcomes

Comparison: logit vs. probit vs. clog-log



The clog-log link function for binary outcomes

Practical considerations:

- Generally, the choice of a link function for binary outcomes is more heavily driven by the interpretation you seek to get out of the model (and less by trying to fit the data perfectly).

Table of Contents

Conditional logistic regression

Setup: 1:1 matched case control data (binary exposure)

- Suppose each case ($Y = 1$) is matched to a control ($Y = 0$).
- Let $m = 1, \ldots, M$ mark the case-control stratum and let $j = 0, 1$ index the individual in the stratum; consider the model:

$$\text{logit}(\text{P}(Y_{jm} = 1 | X_{jm} = x_{jm})) = \alpha_m + x_{jm}\beta$$

- Sample size: $N = 2M$.
- Number of model parameters: $K = M + 1$.
- Maximum likelihood estimate of β is biased (can show $\widehat{\beta} \longrightarrow_p 2\beta$, so that $\widehat{\text{OR}} \longrightarrow_p \text{OR}^2$).

Example: Simulation

```
## Set seed for reproducibility
set.seed(7345)

## Important function
expit <- function(x) {exp(x)/(1 + exp(x))}

## Set number of simulations
nsim <- 1000

## Number of strata
M <- 500

## Set value of parameter of interes
beta <- 1

## Create place to store results
res <- matrix(0, nrow = nsim, ncol = 1)
```

Example: Simulation

```
for (j in 1:nsim) {
  # Case/control outcome
  y <- rep(c(0,1), M)

  ## Matched pair number
  set <- rep(c(1:M), each = 2)

  # Pair-specific parameters
  alpha <- rep(rnorm(M), each = 2)

  ## Binary exposure
  x <- rbinom(2*M, 1, expit(alpha + beta*y))

  ## Fit logistic regression model and extract results
  zz <- glm(y ~ factor(set) +  x, family = binomial(link = "logit"))
  res[j,1] <- as.numeric(coef(zz)[M + 1])
}

## Result
> mean(res)
[1] 2.006405
```

Conditional logistic regression

Brief background: Conditional likelihood

- Suppose you are interested in estimating $\boldsymbol{\beta}$ under the likelihood $f(\mathbf{y}; \boldsymbol{\alpha}, \boldsymbol{\beta})$, where $\boldsymbol{\alpha}$ marks nuisance parameters.
- It is often possible to factor the density in the following useful way:

$$f(\mathbf{w}, \mathbf{v}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{v}; \boldsymbol{\alpha}, \boldsymbol{\beta}) f(\mathbf{w}|\mathbf{v}; \boldsymbol{\beta})$$

  where $\mathbf{V}$ and $\mathbf{W}$ denote sufficient statistics for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.
- The density in blue is the conditional likelihood for $\boldsymbol{\beta}$ under this factorization (**which is not always possible**). When it is possible to factor in this way, the conditional likelihood (which depends upon $\boldsymbol{\beta}$ but not $\boldsymbol{\alpha}$) can be used to estimate $\boldsymbol{\beta}$.

Conditional logistic regression

Brief background: Conditional likelihood in exponential families

- Suppose the density follows the following exponential form:

$$f(\mathbf{y};\boldsymbol{\alpha},\boldsymbol{\beta}) = h(\mathbf{y})\exp\left(\sum_i \beta_i \mathbf{w}_i + \sum_j \alpha_j \mathbf{v}_j - a(\boldsymbol{\alpha},\boldsymbol{\beta})\right).$$

- Then, it can be shown that the conditional distribution of $\mathbf{w}$ (given $\mathbf{v}$) is of the following form:

$$f(\mathbf{w}|\mathbf{v};\boldsymbol{\beta}) = q(\mathbf{w},\mathbf{v})\exp\left(\sum_i \beta_i \mathbf{w}_i - c(\mathbf{v},\boldsymbol{\beta})\right).$$

- Note in particular that the conditional likelihood depends only upon $\boldsymbol{\beta}$ and it can be used for inference.

Conditional logistic regression

Setup: Matched case control data

- Let $i = 1, \ldots, N$ index study units.
- Let $I_{mi} = 1(\text{set}_i = m)$ indicate membership to the m^{th} matched set (of which there are M).
- Let $\mathbf{z}_i = (I_{1i}, \cdots, I_{Mi}, \mathbf{x}_i)$.
- Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \beta) = (\alpha_1, \ldots, \alpha_M, \boldsymbol{\beta})$.
- Then, the likelihood is given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{Z}, \mathbf{y}) &= \prod_{i=1}^{N} \left[\frac{\exp(\mathbf{z}_i^T \boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i^T \boldsymbol{\theta})} \right]^{y_i} \left[1 - \frac{\exp(\mathbf{z}_i^T \boldsymbol{\theta})}{1 + \exp(\mathbf{z}_i^T \boldsymbol{\theta})} \right]^{1-y_i} \\ &= \frac{\exp\left(\sum_{i=1}^{N} \mathbf{z}_i^T \boldsymbol{\theta} y_i \right)}{\prod_{i=1}^{N} (1 + \exp(\mathbf{z}_i^T \boldsymbol{\theta}))}. \end{aligned}$$

Conditional logistic regression

Conditional likelihood: Matched case control data

- Suppose there are M matched sets, each with one case matched to K_m controls. Without loss of generality, let the first observation in each matched set (of size $K_m + 1$) be the case, so that $Y_{1,m} = 1$ and $Y_{2,m} = \cdots = Y_{K_m+1,k} = 0$.
- It can be shown (details omitted) that the conditional likelihood takes the following form:

$$\mathcal{L}_C(\boldsymbol{\beta}; \mathbf{X}) = \prod_{m=1}^{M} \frac{\exp(\mathbf{x}_{1m}^T \boldsymbol{\beta})}{\sum_{k=1}^{K_m+1} \exp(\mathbf{x}_{km}^T \boldsymbol{\beta})}.$$

Conditional logistic regression

Intuition:

- Consider the first stratum of a matched case-control study; say it has one case and two controls. Further, to simplify things, say there is a single dichotomous exposure variable. Label the subjects as $i = 1, 2, 3$.
- Let

$$X_{Di} = \begin{cases} 1 & \text{if subject } i \text{ is case} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad X_{Ei} = \begin{cases} 1 & \text{if subject } i \text{ is exposed} \\ 0 & \text{otherwise} \end{cases}.$$

- If the data consisted of only this stratum, we would maximize the likelihood corresponding to $\text{logit}(p) = \alpha + \beta x_E$ (α is the stratum-specific intercept).

Intuition:

- Make-believe data:

i	x_{Di}	x_{Ei}
1	1	1
2	0	1
3	0	0

- The unconditional likelihood is of the form:

$$\mathrm{P}(X_{D1}=1, X_{D2}=0, X_{D3}=0 | X_{E1}=1, X_{E2}=1, X_{E3}=0; \alpha, \beta),$$

  and we can find the values of α and β that maximize its value.

- Note that the likelihood depends upon the underlying disease risk (case probability) in this stratum (α) and the association between the exposure and disease incidence (β).

Conditional logistic regression

Intuition:

- The unconditional likelihood can be realized as:

$$\text{P(Exactly one is diseased } \textbf{and} \text{ first is diseased}|X_{E1}=1, X_{E2}=1, X_{E3}=0; \alpha, \beta),$$

- The "exactly one subject is diseased" component of the statement is the stratum-specific risk (probability of being a case).
- The "first subject is diseased" component of the statement is the association between the exposure and risk of disease.

Conditional logistic regression

Intuition:

- The conditional likelihood can be realized as:

$$\begin{aligned}&\mathrm{P}(X_{D1}=1, X_{D2}=0, X_{D3}=0 | X_{E1}=1, X_{E2}=1, X_{E3}=0, X_{D1}+X_{D2}+X_{D3}=1; \beta)\\ &\quad = \mathrm{P}(\text{First is diseased} | X_{E1}=1, X_{E2}=1, X_{E3}=0, \text{Exactly one is diseased}; \beta)\end{aligned}$$

- The stratum-specific risk (probability of being a case) is conditioned upon in the conditional likelihood; the association between the risk and exposure does not depend upon the underlying disease risk.
- Since this likelihood does not depend upon α, it does not allow us to estimate α.

Conditional logistic regression

Intuitive exploration: Key ideas

- Conditional logistic regression provides estimates of the coefficient of any variable that varies within at least one stratum.
  - For example, the exposure.
  - For example, confounders that vary within one more strata.
  - Interactions of the above two with "stratum determinants" (i.e., matching variables).
- Conditional logistic regression cannot provide estimates for the coefficient of any variable that does not vary within at least one stratum (discordant pairs).
  - For example, a stratum determinant.
  - For example, interactions between stratum determinants.

Conditional logistic regression

Intuitive exploration: Key ideas

- Model: $\text{logit}(\text{P}(Y_{jm} = 1 | X_{jm} = x_{jm})) = \alpha_m + \mathbf{x}_{jm}^T \beta$.
- Even though they don't appear in the computer output of a conditional logistic regression command, the stratum indicators are still "in the model."
- We impose no constraints on how the underlying probabilities vary from stratum to stratum when we perform conditional logistic regression, *including* their interactions with the exposures.
- Therefore, including interactions between $\mathbf{x}$ and stratum determinants does not violate our usual rules about how main effects need to be included in models that contain their interactions.

Example: Simulation

```
## Important library
library(survival)

## Important function
expit <- function(x) {exp(x)/(1 + exp(x))}

## Set seed for reproducibility
set.seed(7345)

## Set number of simulations
nsim <- 1000

## Number of strata
M <- 500

## Set value of parameter of interes
beta <- 1

## Create place to store results
res <- matrix(0, nrow = nsim, ncol = 1)
```

Example: Simulation

```
for (j in 1:nsim) {
  # Case/control outcome
  y <- rep(c(0,1), M)

  ## Matched pair number
  set <- rep(c(1:M), each = 2)

  # Pair-specific parameters
  alpha <- rep(rnorm(M), each = 2)

  ## Binary exposure
  x <- rbinom(2*M, 1, expit(alpha + beta*y))

  ## Fit logistic regression model and extract results
  zz <- clogit(y ~ x + strata(set))
  res[j,1] <- as.numeric(coef(zz))
}

## Result (WHEW!!)
> mean(res)
[1] 1.003203
```

Conditional logistic regression

Further thoughts:

- Conditional logistic regression allows you to adjust for confounders even if you didn't match on them (or matched loosely).
- Conditional logistic regression is useful because it easily accounts for matched sets with differing numbers of cases and controls (under an unconditional likelihood, this would be more challenging).
- Conditional logistic regression is useful not *only* for matched data but for setting in which you have a lot of adjustment variables that are not of primary scientific interest.
- I expect you will discuss conditional and marginal likelihoods in much grater depth in BIOS 7346.

Summary

So far:

- Additional considerations for binary outcomes.

Summary

Up next:

- Receiver operating characteristic regression.