# Binary Data

Suppose $Y$ is a binary response variable (taking values 0, 1).
$X_1, \ldots, X_p$ are the explanatory variables.
**Goal:**
model the relation between

$$\mathbb{E}(Y) = \pi = \Pr(Y = 1)$$

and $X_1, \ldots, X_p$, based on $n$ independent samples.

# Clinical Trial Example

Binary response variable $Y$: indicator of responder. Predictors $(X_1, ..., X_p)$: Treatment group, Gender, etc.

Ungrouped data $(y_i, \boldsymbol{x}_{(i)})$:

```
Response    Treatment    Gender
---------------------------------
   y(1)         A         male
   ...
   y(m1)        A         male
---------------------------------
   y(m1+1)      B         male
   ...
   y(m1+m2)     B         male
---------------------------------
   y(m1+m2+1)   A       female
   ...
   y(m1+m2+m3)  A       female
---------------------------------
   ...
```

Grouped representation (categorical predictors):

| Treatment | Gender | Group size | # of responders |
|---|---|---|---|
| A | male | m1 | Z1 |
| A | female | m2 | Z2 |
| B | male | m3 | Z3 |
| B | female | m4 | Z4 |

- Data: $(m_i, Z_i, \boldsymbol{x}_{(i)})$
- Even if grouped, individuals in the same group are assumed to be independent, and have the same probability of event occurrence.
- Estimates from the grouped and ungrouped (sparse) data are the same.
- If there are unique values (e.g., age) for each individual, data cannot be grouped.
- Note: there are differences in model diagnostics!

# Binomial Distribution

For binary response $Y \in \{0, 1\}$, the only distribution is the Bernoulli distribution.

- Bernoulli Distribution $Bin(1, \pi)$. $\mathbb{P}(Y = 1) = \pi$, $\mathbb{P}(Y = 0) = 1 - \pi$; or

$$\mathbb{P}(Y = y) = \pi^y (1 - \pi)^{1-y}, \ y = 0, 1.$$

If $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} Bin(1, \pi)$, then $Y = \sum_{i=1}^{n} Y_i$ has a Binomial distribution.

- Binomial Distribution: $Bin(n, \pi)$

$$\mathbb{P}(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \ y = 0, \ldots, n.$$

# Models for Binary data

Suppose $Y_i \sim Bin(1, \pi_i)$; $X_i$ is a $p \times 1$ covariate vector for individual $i$; $\eta_i = X_i \beta$ is the linear predictor. The log-likelihood function is

$$\sum_{i=1}^{n} \left[ Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i) \right] \triangleq \sum_{i=1}^{n} \left[ Y_i \theta_i - b(\theta_i) \right].$$

where

$$\theta_i = \log \frac{\pi_i}{1 - \pi_i}, \;\; b(\theta_i) = \log(1 + \exp(\theta_i))$$

The last piece of GLM is to link $\pi_i$ with $\eta_i$, using some link function $g(\cdot)$

$$g(\pi_i) = \eta_i = X_i^T \beta$$

# Models for Grouped Data

Suppose $(Y_i, n_i, X_i)$ is the observed data, where $Y_i \sim Bin(n_i, \pi_i)$. The log likelihood function is

$$\sum_{i=1}^{m} \left\{ \log \binom{n_i}{Y_i} + [Y_i \log \pi_i + (n_i - Y_i) \log(1 - \pi_i)] \right\}$$

where

$$\theta_i = \log \frac{\pi_i}{1 - \pi_i}, \;\; b(\theta_i) = n_i \log(1 + \exp(\theta_i))$$

Again, we need to define a link function $g(\cdot)$ to link the mean $n_i \pi_i$ and $\eta_i$.

# Link functions

$$0 \le \pi_i \le 1,$$

$$\eta_i = X_i^T \beta \in \mathbb{R},$$

which suggests that a link function must satisfy

$$g : [0, 1] \mapsto \mathbb{R},$$

$$g(0) = -\infty, \text{ and } g(1) = +\infty.$$

Correspondingly, $g^{-1} : \mathbb{R} \mapsto [0, 1]$ and monotone increasing.

1. **logit/logistic (the canonical link)**:

$$g_1(\pi) = {b'}^{-1}(\pi) = \log \frac{\pi}{1-\pi},$$

$$g_1^{-1}(\eta) = \frac{e^\eta}{1+e^\eta}.$$

2. probit/inverse Normal:

$$g_2(\pi) = \Phi^{-1}(\pi),$$

$$g_2^{-1}(\eta) = \Phi(\eta).$$

3. complementary log-log:

$$g_3(\pi) = \log(-\log(1-\pi)),$$

$$g_3^{-1}(\eta) = 1 - e^{-e^\eta}.$$

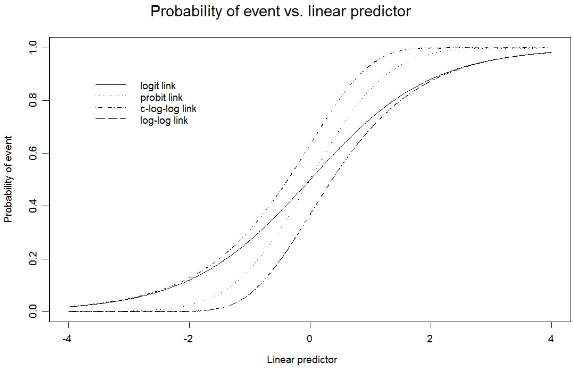


Figure: The choice of link functions depend on model fit and interpretation!

- $g_1(\pi) = -g_1(1-\pi)$, $g_2(\pi) = -g_2(1-\pi)$
- $g_3(\pi)$ is not symmetric.

# Parameter Interpretation

Link function determines how the linear predictor exert its effect on the event rate/risk.
Consider a logistic model with two covariates $X_1$ and $X_2$:

$$\log(\frac{\pi}{1-\pi}) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

- $\pi$ is usually called the **risk**
- $\pi/(1-\pi)$ is called the **odds**
- $\beta_0$ is the log odds for $X_1 = X_2 = 0$;
- $\beta_1$ is the unit change in log odds (or **log odds ratio**) per unit change of $X_1$ holding $X_2$ fixed (assuming $X_1$, $X_2$ independent).
- if $\beta_1 > 0$, the risk increases with $X_1$; vice versa.

# Examples

- **Show/no-show:**
  Investigating the relation between show/no-show and appointment lag.
- **Peer reviewed publication:**
  Comparing urology fellows with and without time off in terms of their proportions of urological publications.