Lecture 6
GLM for binary data: introduction

Today’s topics:

- Binary / Binomial data model
  - Data input
  - Link functions
  - R example
- Reading: Agresti Chapter 5.1,5.6, Faraway Chapters 2.1, 3.1, 4.1-4.2

Binary / binomial data model

If the observation y_i is binomial

$$y_i \sim \text{Binomial}(n_i, p_i)$$

and probability function:

$$f(y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} = \binom{n_i}{y_i} \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i)^{n_i}$$

If $n_i = 1$, then y_i is a 0/1 binary data point (follows a Bernoulli distribution).

- Link function: $g(p_i) = X_i^T \beta$
- logistic regression: $\log(\frac{p_i}{1-p_i}) = X_i^T \beta$

Data input for binary model

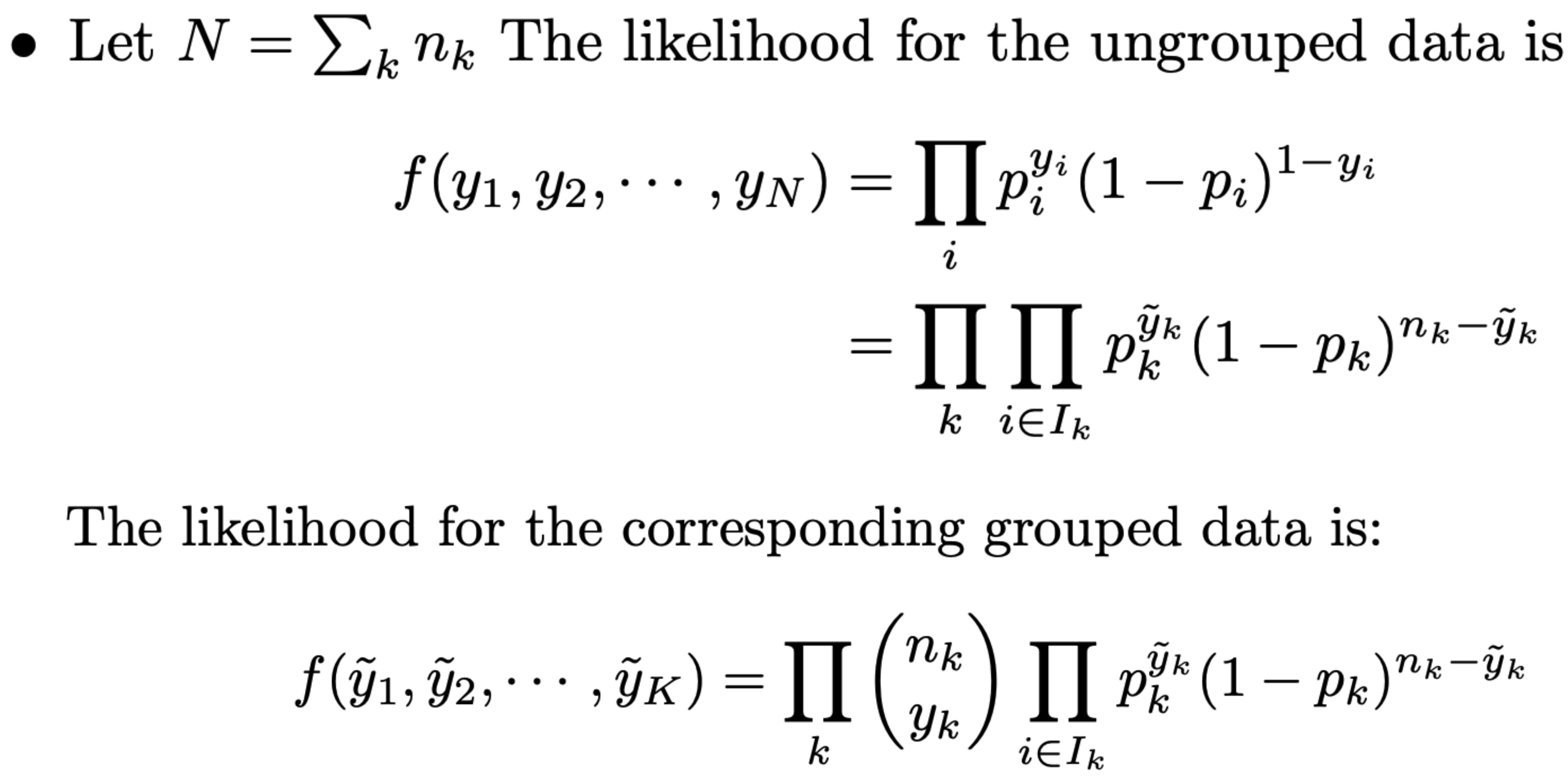
If X_i are categorical variables, then we may have samples with the same Xi and we can group them together

- ungrouped data: each $n_i = 1$ and some samples have the same X_i, thus they share the same p_i
- a grouped sample $\tilde{y}_k$ for group k where all observations in the group share the same X_i
  - Define n_k as the number of binary observations
  - The grouped response for group k is

$$\tilde{y}_k = \sum_{i \in I_k} y_i \sim \text{Binomial}(n_k, p_k)$$

- The grouped data follows the Binomial distribution because we assume that the samples are independent within each group

Likelihood for grouped and ungrouped data



- Let $N = \sum_k n_k$ The likelihood for the ungrouped data is:

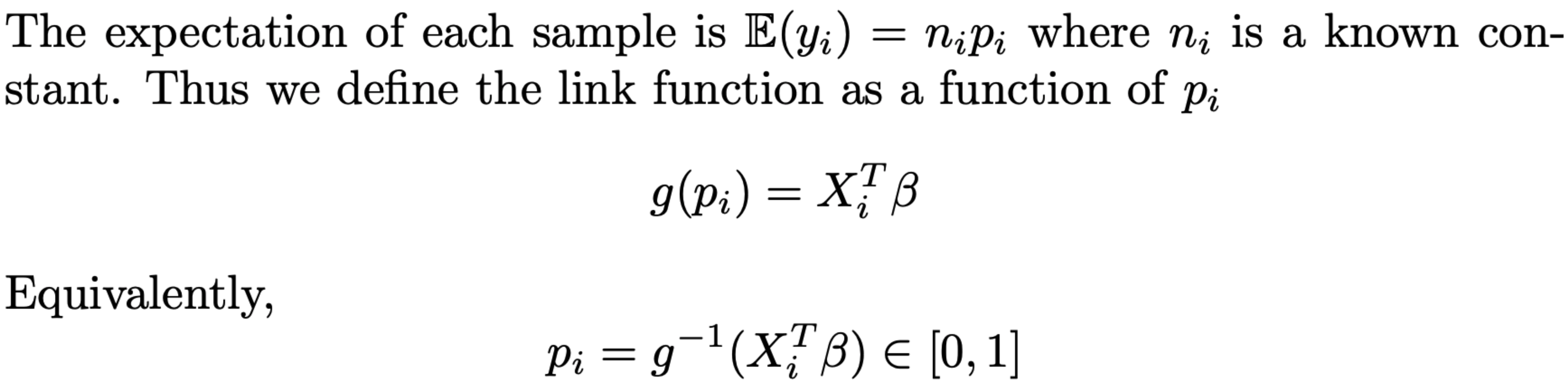
$$f(y_1, y_2, \cdots, y_N) = \prod_i p_i^{y_i}(1-p_i)^{1-y_i}$$
$$= \prod_k \prod_{i \in I_k} p_k^{\tilde{y}_k}(1-p_k)^{n_k-\tilde{y}_k}$$

The likelihood for the corresponding grouped data is:

$$f(\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_K) = \prod_k \binom{n_k}{y_k} \prod_{i \in I_k} p_k^{\tilde{y}_k}(1-p_k)^{n_k-\tilde{y}_k}$$

- The likelihood is not the same between the grouped data and ungrouped data. However, the log-likelihood function only differs by a constant, thus the GLM solution does not change.

Link function for binary / binomial GLM



The expectation of each sample is $\mathbb{E}(y_i) = n_i p_i$ where n_i is a known constant. Thus we define the link function as a function of p_i

$$g(p_i) = X_i^T \beta$$

Equivalently,

$$p_i = g^{-1}(X_i^T \beta) \in [0, 1]$$

- If g is a one-to-one mapping and continuous function, then g^{-1} should be monotone.
- one natural choice of g^{-1} is to make it as a cdf of some distribution.
- Denote $F(z) = g^{-1}(z)$ as some cdf function
  - Let $\epsilon_i \overset{i.i.d.}{\sim} F(\cdot)$

Latent variable threshold models

- Denote $F(z) = g^{-1}(z)$ as some cdf function
  - Let $\epsilon_i \overset{i.i.d.}{\sim} F(\cdot)$
  - Then

$$p_i = F(X_i^T \beta) = \mathbb{P}(\epsilon_i \leq X_i^T \beta) = \mathbb{P}\left(X_i^T \beta - \epsilon_i >= 0\right)$$

  - This is called a latent variable threshold models and $X_i^T \beta - \epsilon_i$ are the “latent variables”

  - It does not make any essential modeling difference choosing the cutoff to be 0 or any other value τ

Latent variable threshold models

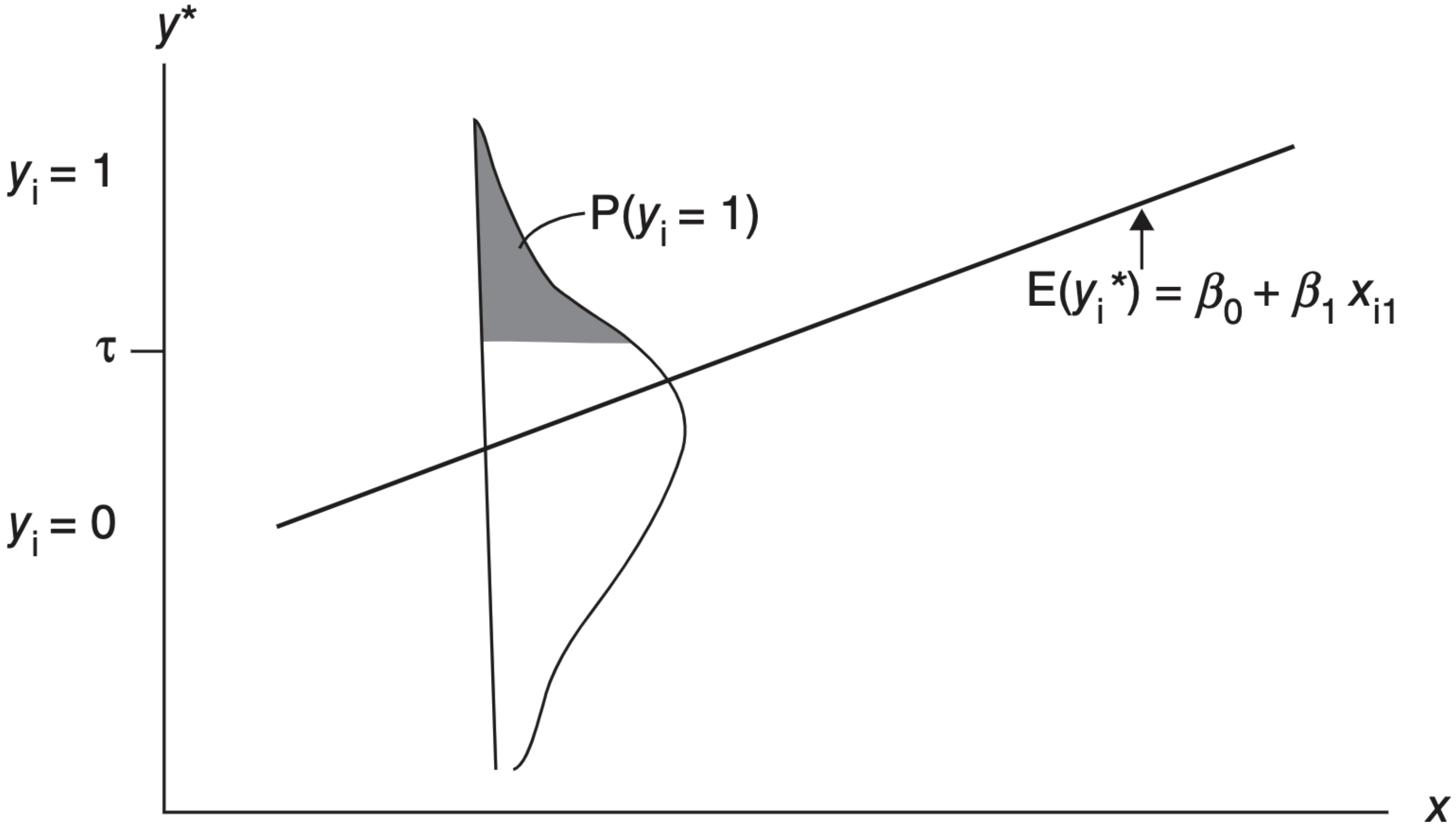


Figure 5.1 Threshold latent variable model, for which we observe $y_i = 1$ when underlying latent variable $y_i^* > \tau$.

The probit link

- The probit link: $F(z)$ is the cdf of a standard Gaussian distribution

$$p_i = \mathbb{P}\left(X_i^T\beta - \epsilon_i >= 0\right) = \mathbb{P}\left(X_i^T\beta + \epsilon_i >= 0\right)$$

where $\epsilon_i \sim N(0,1)$. Let the hidden variable be $y_i^\star = X_i^T\beta + \epsilon_i$, then it goes to the definition of the probit link that some of you may be more familiar with:

$$Y_i = \begin{cases} 1 & \text{if } y_i^\star >= 0 \\ 0 & \text{else} \end{cases}$$

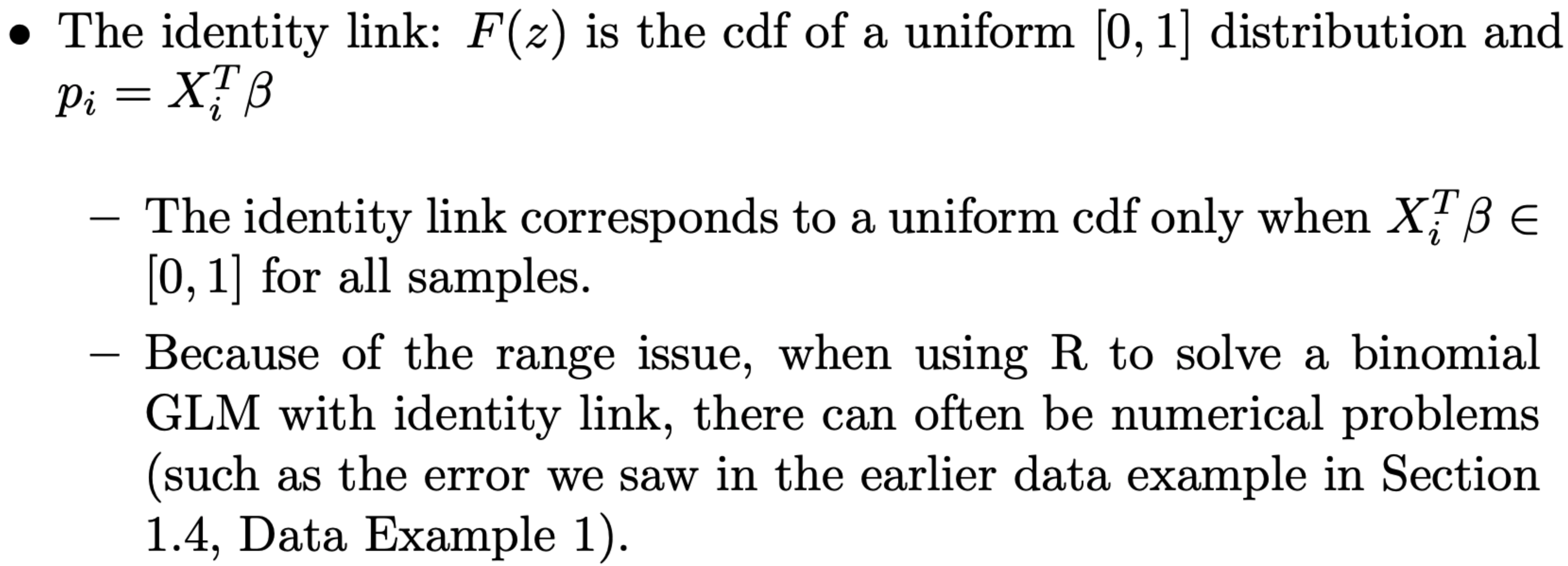
The logit link

- The logit link: $F(z)$ is the cdf of a standard logistic distribution

$$F(z) = \frac{e^z}{1+e^z}$$

  - The link function is called the logit link: $g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$
  - The logit link is the canonical link of the Binomial distribution

The identity link



- The identity link: $F(z)$ is the cdf of a uniform $[0,1]$ distribution and $p_i = X_i^T\beta$
  - The identity link corresponds to a uniform cdf only when $X_i^T\beta \in [0,1]$ for all samples.
  - Because of the range issue, when using R to solve a binomial GLM with identity link, there can often be numerical problems (such as the error we saw in the earlier data example in Section 1.4, Data Example 1).

The log-log link

- All previous links assume a symmetric ϵ_i around 0:

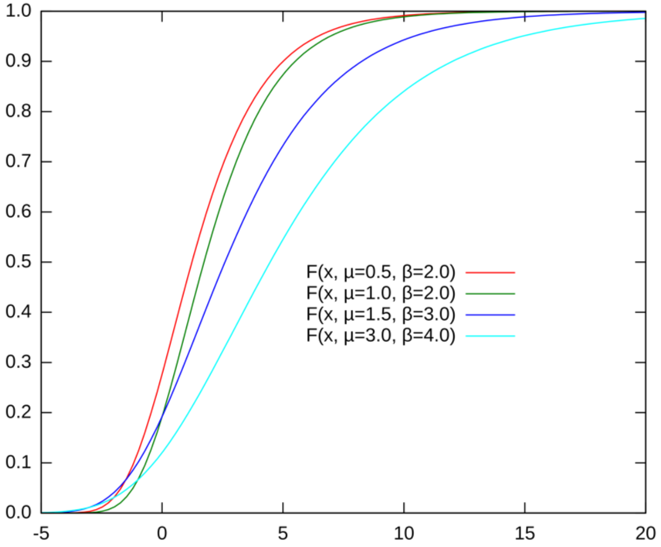
$$F(x) = 1 - F(-x) \Leftrightarrow g(p) = -g(1-p)$$

  - A corresponding restriction is that the response curve is symmetric at 0.5
  - We should use some other link functions (or F) if this assumption is severely violated
    - The log-log link: $F(z)$ is the cdf of a standard double-exponential distribution (Gumbel distribution)

$$F(z) = e^{-e^{-z}}$$

– The link function is called the log-log link:

$$g(p_i) = -\log[-\log(p_i)] = X_i^T\beta$$



  - With the log-log link, p_i approaches 0 sharply but approaches 1 slowly

The complementary log-log link

- With a complementary log-log link, p_i approaches 1 sharply but approaches 0 slowly

$$g(p_i) = \log[-\log(p_i)] = X_i^T \beta$$

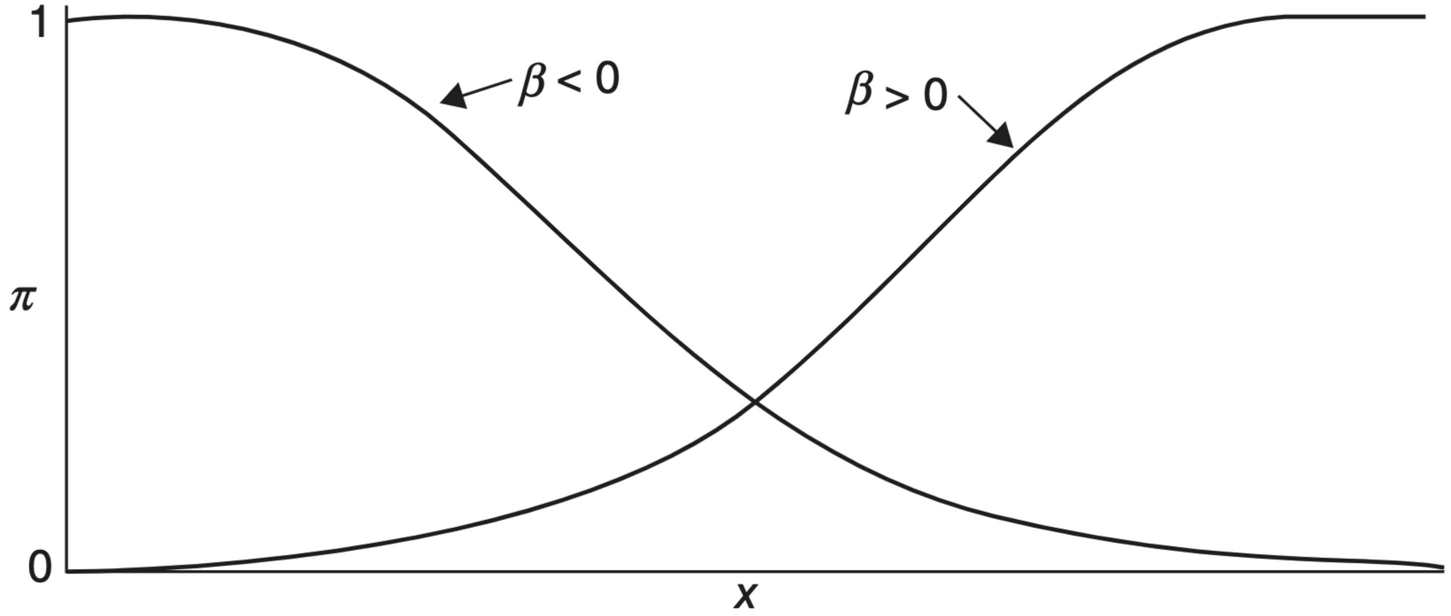


Figure 5.4 GLM for binary data using complementary log–log link function.

R data example for binary / binomial GLM (part I)

- Check Example3_1 R notebook