Recall that we specified a glm for binary data as

1. $Y_i \sim \text{Bernoulli}(\pi_i)$ independent $i = 1, \dots, n$

   hence $\pi_i = \mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1)$, $\pi_i \in [0,1]$
2. $\eta_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} = \tilde{x}_i^T \beta$
3. $g(\pi_i) = \eta_i$

We analyzed the case where $g(\cdot)$ is the canonical link function: logit model

However, g could be any function that maps $[0,1] \to \mathbb{R}$, invertible (and differentiable).

$\rightarrow$ (inverse of) cumulative distribution functions are good candidates.

- **INTERPRETATION AS THRESHOLD MODEL**

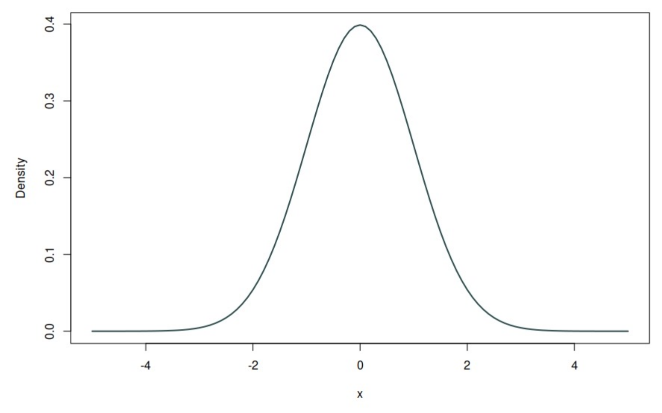
Assume that $Y_i \sim \text{Bernoulli}(\pi_i)$ $i = 1, \dots, n$ and

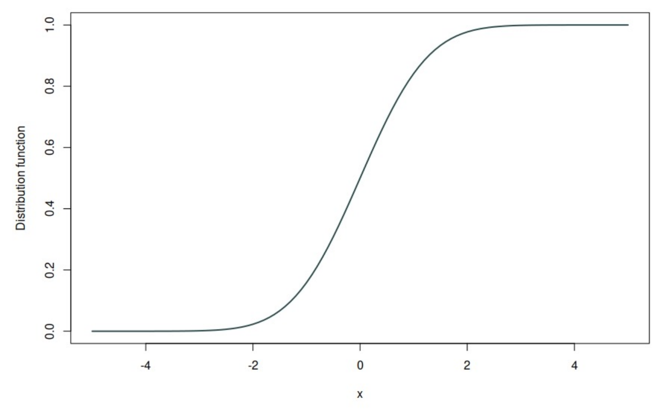
$\pi_i = F(\tilde{x}_i^T \beta)$ with F the cumulative distribution function of a random variable with distribution SYMMETRIC around zero

Then the regression for Y_i has an interpretation in terms of a regression model on a CONTINUOUS LATENT (= unobserved) random variable z

Let us consider, for example, the PROBIT REGRESSION MODEL

Here, $F = \Phi$ is the CDF of a standard Gaussian distribution





PROBIT REGRESSION: assumptions

- $Y_i \sim \text{Bern}(\pi_i)$ indep. for $i = 1, \dots, n$
- $\eta_i = \tilde{x}_i^T \beta$ linear predictor
- $g(\pi_i) = \Phi^{-1}(\pi_i) = \eta_i$ with Φ^{-1} quantile function of a $N(0,1)$

  $\Rightarrow$ we obtain $\pi_i = \Phi(\tilde{x}_i^T \beta)$

Example: study on a treatment for hypertension (high blood pressure)

We observe a binary response variable

$$Y_i = \begin{cases} 1 & \text{if subject } i \text{ has hypertension} \\ 0 & \text{if subject } i \text{ does not have hypertension} \end{cases}$$

we can only observe this binary version, but actually there is an underlying continuous r.v. (that we do not have)

z_i = blood pressure (mmHg)

We can think of Y_i as a "simplified" measure, obtained starting from Y_i^*:

$$Y_i = \begin{cases} 1 & \text{if } z_i > k \\ 0 & \text{if } z_i \leq k \end{cases} \qquad k = \text{threshold (fixed)}$$

In the example:

Subject i has hypertension ($y_i = 1$) if their blood pressure is above 140/90 mmHg.

Model:

For simplicity, we assume $k = 0$. When the threshold is $k \neq 0$, it is sufficient to consider as latent random variable $(z_i - k)$

We assume a GAUSSIAN LINEAR MODEL on the LATENT VARIABLE z_i

Assumptions:

$$\left.\begin{array}{ll} z_i = \tilde{x}_i^T \beta + \varepsilon_i & i = 1, \dots, n \\ \varepsilon_i \text{ iid with distribution } & \varepsilon_i \sim N(0, 1) \end{array}\right\} \Rightarrow z_i \sim N(\tilde{x}_i^T \beta, 1) \text{ indep. } i = 1, \dots, n$$

we assume known variance $= 1$

However, we do not have z_i, but only its dichotomized version Y_i:

$$Y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0 \end{cases}$$

what is $\mathbb{P}(Y_i = 1) = \pi_i$?

$$\begin{aligned} \mathbb{P}(Y_i = 1) &= \mathbb{P}(z_i > 0) \\ &= 1 - \mathbb{P}(z_i \leq 0) \\ &= 1 - \mathbb{P}(\tilde{x}_i^T \beta + \varepsilon_i \leq 0) \\ &= 1 - \mathbb{P}(\varepsilon_i \leq -\tilde{x}_i^T \beta) \\ &= 1 - \Phi(-\tilde{x}_i^T \beta) \\ &= 1 - (1 - \Phi(\tilde{x}_i^T \beta)) = \Phi(\tilde{x}_i^T \beta) \end{aligned}$$

$\Rightarrow \pi_i = \Phi(\tilde{x}_i^T \beta)$

which is exactly the model we assumed for Y_i (GLM).

Probit regression can be interpreted as a "simplification" of a Gaussian linear model, where we do not have all information on z_i but only a dichotomized version.