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).

$\rightarrow$ cumulative distribution functions are good candidates.

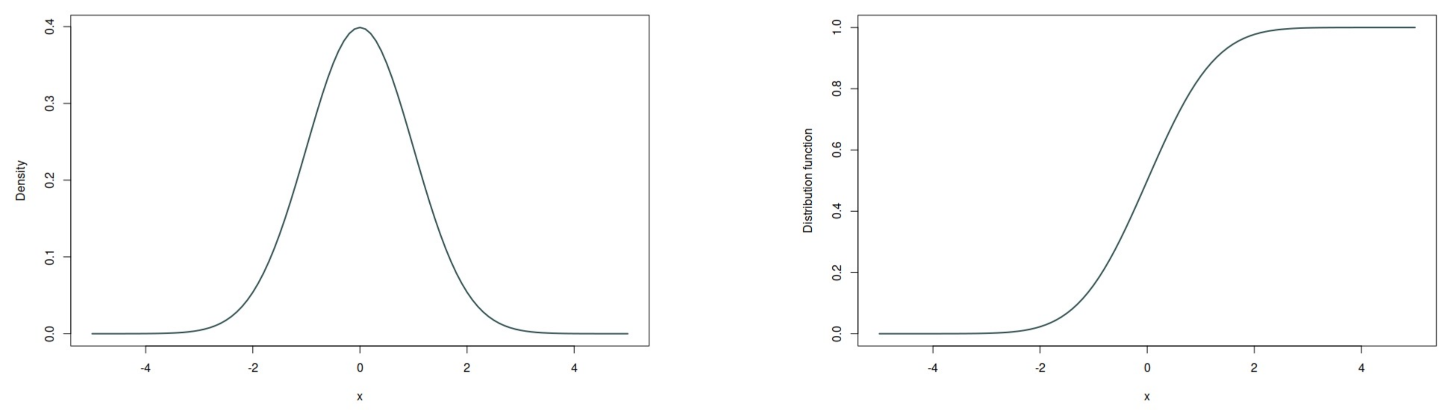
- INTERPRETATION as THRESHOLD MODEL

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 cdf of a r.v. with distribution SYMMETRIC around zero

Then the regression for Y_i has an interpretation in terms of a model on a CONTINUOUS LATENT (= unobserved) r.v. Y_i^*.

Let us consider, for example, the PROBIT MODEL, where $F = \Phi$ is the cumulative distribution function of a standard Gaussian distribution :



PROBIT REGRESSION : model assumptions

- $Y_i \sim \text{Bernoulli}(\pi_i)$ $i = 1, \dots, n$ independent
- $\eta_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} = \tilde{x}_i^T \beta$
- $g(\pi_i) = \Phi^{-1}(\pi_i) = \eta_i$

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

example : Y_i = subject i has high blood pressure = $\begin{cases} 1 & \text{hypertension} \\ 0 & \text{no hypertension} \end{cases}$

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

Indeed we can assume that Y_i is obtained starting from Y_i^* as

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > k \\ 0 & \text{if } Y_i^* \le k \end{cases}$$

" subject i is considered to have high blood pressure if their pressure is above a threshold k "

$\rightarrow$ For simplicity, we consider $k = 0$ (it is sufficient to consider $Y_i^* - k$ for $k \neq 0$)

We assume a GAUSSIAN LINEAR MODEL on the LATENT VARIABLE Y_i^*

$$\left.\begin{aligned} & Y_i^* = \tilde{x}_i^T \beta + \varepsilon_i \quad i = 1, \dots, n \\ & \varepsilon_i \text{ iid with distribution } \varepsilon_i \sim N(0, \underbrace{1}_{\text{known variance} = 1}) \end{aligned}\right\} \Rightarrow Y_i^* \sim N(\tilde{x}_i^T \beta, 1) \text{ independent}$$

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

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \le 0 \end{cases}$$

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

$$\begin{aligned} \mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i^* > 0) &= 1 - \mathbb{P}(Y_i^* \le 0) = 1 - \mathbb{P}(\tilde{x}_i^T \beta + \varepsilon_i \le 0) = \\ &= 1 - \mathbb{P}(\varepsilon_i \le -\tilde{x}_i^T \beta) \qquad \varepsilon_i \sim N(0,1) \\ &= 1 - \Phi(-\tilde{x}_i^T \beta) \\ &= \Phi(\tilde{x}_i^T \beta) \end{aligned}$$

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

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 Y_i^* but only a dichotomized version.