BINARY REGRESSION

The response variable Y_i is a binary r.v. (takes only 2 values).
This kind of setting can be found in various studies: the 2 states can represent, e.g., presence / absence, success / failure, alive / dead, ...

The data can be organized into 2 different forms:
1. UNGROUPED: the response vector is the output of each statistical unit $\Rightarrow \in \{0,1\}$
2. GROUPED: if for each combination of the covariates I observe several units, it is possible to aggregate the data by summing the # of 1 and 0 for each combination.
(grouped data can be converted to the ungrouped form)

Example: Beetle data. Study on the efficacy of a beetle poison for killing beetles
x_i = (log) dose of poison
unprouped data: $Y_i \in \{0,1\}$ $\quad Y_i = \begin{cases} 0 & \text{if the i-th beetle is alive} \\ 1 & \text{if the i-th beetle is dead} \end{cases}$

dead / alive	0	1	...	0	0	...	1	...	1	...	0	1
log(dose)	1.69	1.69	...	1.69	1.724	...	1.724	.....	1.88	...	1.88	1.88

The experiment has been run on several beetles for every dose: I can count how many beetles are dead or alive for each level. I obtain the grouped data

# killed (1)	6	13	...	60
# alive (0)	53	47	...	0
log(dose)	1.69	1.724	...	1.88

notice: the number of beetles for each level of x_i need not be the same

For the ungrouped data, a reasonable model is the Bernoulli
$Y \sim \text{Bern}(\pi)$
- parameter space: $\pi \in [0,1]$ $\qquad \pi = \mathbb{P}(Y=1)$ is a probability
- support: $\mathcal{Y} = \{0,1\}$
- probability mass function $\quad p(y;\pi) = \mathbb{P}(Y=y) = \pi^y(1-\pi)^{1-y}$
- moments: $\mathbb{E}[Y] = \pi, \quad \text{var}(Y) = \pi(1-\pi)$

For the grouped data, we would use a binomial (more about it later).

- BINARY REGRESSION: ASSUMPTIONS with UNGROUPED DATA
1. $Y_i \sim \text{Bernoulli}(\pi_i)$ independent $i = 1,\dots,n$
   hence $\pi_i = \mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1), \quad \pi_i \in [0,1]$
2. $\eta_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} = \underset{\sim}{x}_i^T \underset{\sim}{\beta}$
3. $g(\pi_i) = \eta_i$

Remarks:
- LINK FUNCTION
GLMs model the MEAN of the random variables: here $\mathbb{E}[Y_i] = \pi_i$, which is also $\mathbb{P}(Y_i=1)$.
π_i is a probability $\Rightarrow \pi_i \in [0,1]$. However, $\eta_i \in \mathbb{R} \rightarrow g$ should be a function that maps $[0,1] \rightarrow \mathbb{R}$, invertible (and differentiable). For simplicity, it is usually assumed monotone increasing. Common choices are
  - $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ LOGISTIC function (hence the "logistic regression": this is the canonical link)
    (inverse of the distribution function of the logistic distribution)
  - $g(\pi_i) = \Phi^{-1}(\pi_i)$ PROBIT function, where Φ is the distribution function of a Gaussian distr.

- VARIANCE
The Bernoulli distribution assumes $\text{var}(Y_i) = \pi_i(1-\pi_i) = \mathbb{E}[Y_i](1-\mathbb{E}[Y_i])$.
Hence, again, the random variables are not homoscedastic.