28 Nov - LEC 11

BINARY REGRESSION

The response variable Y_i is BINARY (takes only 2 values).
Several experiments have a binary outcome: the 2 states can represent, for example, presence / absence, success / failure, alive / dead, ...
The two states are encoded with the values 0 and 1 $\Rightarrow$ $y_i \in \{0,1\}$

Similar to previous settings, for each unit we also observe p covariates $(x_{i1}, \dots, x_{ip})$
$\rightarrow$ data: $y_i \in \{0,1\}$ and $(x_{i1}, x_{i2}, \dots, x_{ip})$ for $i = 1, \dots, n$.

The data can be organized in two different ways:

1. UNGROUPED: each element of the response vector is the realisation of an individual experiment $y_i \in \{0,1\}$
2. GROUPED: IF FOR SOME COMBINATIONS OF COVARIATES I OBSERVE SEVERAL UNITS, it is possible to aggregate these outcomes by COUNTING THE NUMBER OF 0 AND 1 FOR EACH COMBINATION.

(Notice that all grouped data can be converted to the ungrouped form; however, the contrary is not true since we don't always have several units with equal covariates).
Let's say we have m units with all covariates equal. Grouped data then count:

$$\sum_{i=1}^{m} y_i = \text{number of ones}$$
$$m - \sum_{i=1}^{m} y_i = \text{number of zeros}$$

Example: Beetle data. Study on the efficacy of a beetle poison for killing beetles

- x_i = log-dose of poison
- outcome = beetle i is dead / alive
  ↳ we encode the outcome as $y_i = \begin{cases} 1 & \text{if beetle } i \text{ is dead} \\ 0 & \text{if beetle } i \text{ is alive} \end{cases}$

- UNGROUPED DATA

each $y_i \in \{0,1\}$

y_i	x_i
0	1.69
0	1.69
⋮	⋮
0	1.69
0	1.72
1	1.72
⋮	⋮
0	1.72
⋮	⋮
⋮	⋮
1	1.88
1	1.88
⋮	⋮
1	1.88

each dose is applied to several beetles

- GROUPED DATA

Since the experiment has been repeated on several beetles for each dose of poison, we can count how many beetles are dead or alive at each dose level.
We obtain grouped data:

# dead	# alive	x_i
6	53	1.69
13	47	1.72
⋮	⋮	
60	0	1.88
↓ number of ones	↓ number of zeros	

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: $\pi \in [0,1]$ $\qquad \pi = \mathbb{P}(Y=1)$ success probability
- support: $\mathcal{Y} = \{0,1\}$
- probability mass function $p(y \mid \pi) = \mathbb{P}(Y=y) = \pi^y (1-\pi)^{1-y}$
- moments: $\mathbb{E}[Y] = \pi$, $\quad \text{var}(Y) = \pi(1-\pi)$

For the grouped data, we use a binomial
$Y \sim \text{Bi}(n, \pi)$

- parameter space: $\pi \in [0,1]$ success probability
  $n \in \{0, 1, 2, \dots\}$ number of trials
- support: $\mathcal{Y} = \{0, 1, \dots, n\}$
- probability mass function $p(y; n, \pi) = \mathbb{P}(Y=y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$
- moments: $\mathbb{E}[Y] = n\pi$ $\quad \text{var}(Y) = n\pi(1-\pi)$

BINARY REGRESSION:
general assumptions with UNGROUPED DATA

- distribution $Y_i \sim \text{Bern}(\pi_i)$ independent for $i = 1, \dots, n$
  hence $\mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1) = \pi_i$
- linear predictor $\eta_i = \underset{\sim}{x}_i^T \underline{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$
- link function $g(\pi_i) = \eta_i$

Remark: the 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 $[0,1] \to \mathbb{R}$, invertible (and differentiable).
For simplicity, it is usually assumed monotone INCREASING. Common choices are

- $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ LOGIT FUNCTION (inverse of the CDF of the logistic distribution)
  it is the CANONICAL LINK. We obtain the so called "logistic regression"
- $g(\pi_i) = \Phi^{-1}(\pi_i)$ PROBIT function, where Φ is the distribution function of a Gaussian distr.

Remark: 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.