28 Nov - LEC 11

LOGISTIC REGRESSION for UNGROUPED DATA

MODEL ASSUMPTIONS:

- $Y_i \sim \text{Bern}(\pi_i)$ indep. $i = 1, \dots, n$
- $\eta_i = \underset{\sim}{x}_i^T \underline{\beta}$
- $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$ LOGIT FUNCTION

if we invert the relationship between π_i and η_i we obtain

$$\pi_i = g^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} \in (0,1)$$

Hence we can write the model as

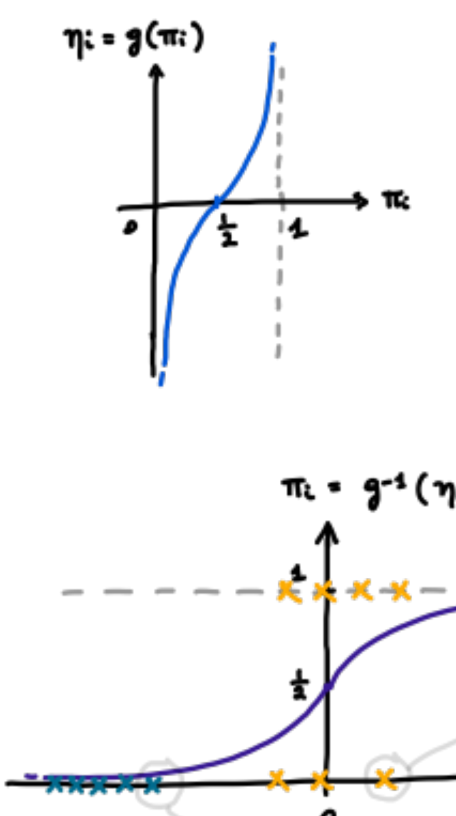
$Y_i \sim \text{Bern}(\pi_i)$ independent for $i = 1, \dots, n$

with $\pi_i = g^{-1}(\underset{\sim}{x}_i^T \underline{\beta}) = \frac{e^{\underset{\sim}{x}_i^T \beta}}{1+e^{\underset{\sim}{x}_i^T \beta}} = \mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1)$

and the distribution of Y_i is

$$\mathbb{P}(Y_i = y_i) = \Bigg(\underbrace{\frac{e^{\underset{\sim}{x}_i^T \beta}}{1+e^{\underset{\sim}{x}_i^T \beta}}}_{\pi_i}\Bigg)^{y_i} \Bigg(\underbrace{\frac{1}{1+e^{\underset{\sim}{x}_i^T \beta}}}_{1-\pi_i}\Bigg)^{1-y_i}$$

REMARK: the logit function



$$\eta_i = \log \frac{\pi_i}{1-\pi_i}$$

if $\pi_i \to 0$, $\text{logit}(\pi_i) \to -\infty$

if $\pi_i \to 1$, $\text{logit}(\pi_i) \to +\infty$

If I imagine to draw Bernoulli samples for different values of η_i:

- if $\eta_i \ll 0 \Rightarrow \pi_i$ close to zero: I observe many failures
- if $\eta_i \approx 0 \Rightarrow \pi_i \approx 0.5$ similar number of failures and successes
- if $\eta_i \gg 0 \Rightarrow \pi_i$ close to 1: many successes

INTERPRETATION OF THE MODEL PARAMETERS

Logistic regression has an interpretation of the parameters in terms of LOG-ODDS

Indeed, $\eta_i = \log \frac{\pi_i}{1-\pi_i}$

The ratio $\frac{\pi_i}{1-\pi_i}$ = ODDS = $\frac{\text{prob. of success}}{\text{prob. of failure}}$

If I consider odds $\cdot\, 100 = \frac{\pi_i}{1-\pi_i} \cdot 100$ → it is the EXPECTED NUMBER OF SUCCESSES EVERY 100 FAILURES

e.g. if odds = 2 in the beetles experiment ⇒ success = dead, fail = alive.
" every 100 alive beetles, 200 are killed "

Notice that odds $= 2 \iff \frac{\pi_i}{1-\pi_i} = 2 \iff \pi_i = \frac{2}{3}$

Since $\log \frac{\pi_i}{1-\pi_i} = \eta_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$ → it is a linear model for the log-odds

The coefficient β_j is the CHANGE IN THE LOG-ODDS IF x_j IS INCREASED BY 1 UNIT, WHILE KEEPING THE OTHER COVARIATES FIXED.

Alternative: we do the usual reasoning

Let's study the mean $\mathbb{E}[Y]$ for two individuals i and k with all the covariates equal except the j-th one, for which we assume $x_{kj} = x_{ij} + 1$

i.e., $x_{ih} = x_{kh}$ for $h = 1, \dots, p$ $h \neq j$, $x_{kj} = x_{ij} + 1$

For individual i we get

$$\mathbb{E}[Y_i] = \pi_i = \frac{e^{\underset{\sim}{x}_i^T \beta}}{1+e^{\underset{\sim}{x}_i^T \beta}} = \frac{\exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \dots + \beta_p x_{ip}\}}{1 + \exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \dots + \beta_p x_{ip}\}}$$

For individual k we get

$$\mathbb{E}[Y_k] = \pi_k = \frac{e^{\underset{\sim}{x}_k^T \beta}}{1+e^{\underset{\sim}{x}_k^T \beta}} = \frac{\exp\{\beta_1 + \beta_2 \overset{x_{i2}}{\overset{\parallel}{x_{k2}}} + \dots + \beta_{j-1} \overset{x_{i,j-1}}{\overset{\parallel}{x_{k,j-1}}} + \beta_j \overset{x_{ij}+1}{\overset{\parallel}{x_{kj}}} + \beta_{j+1} x_{k,j+1} + \dots + \beta_p x_{kp}\}}{1 + \exp\{\beta_1 + \beta_2 \underset{x_{i2}}{\underset{\parallel}{x_{k2}}} + \dots + \beta_{j-1} \underset{x_{i,j-1}}{\underset{\parallel}{x_{k,j-1}}} + \beta_j \underset{x_{ij}+1}{\underset{\parallel}{x_{kj}}} + \beta_{j+1} x_{k,j+1} + \dots + \beta_p x_{kp}\}}$$

The ODDS for individual i:

$$\frac{\pi_i}{1-\pi_i} = \frac{e^{\underset{\sim}{x}_i^T \beta}}{1+e^{\underset{\sim}{x}_i^T \beta}} \cdot \left(\frac{1}{1+e^{\underset{\sim}{x}_i^T \beta}}\right)^{-1} = e^{\underset{\sim}{x}_i^T \beta} = \exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \dots + \beta_p x_{ip}\}$$

The ODDS for individual k:

$$\frac{\pi_k}{1-\pi_k} = e^{\underset{\sim}{x}_k^T \beta} = \exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{i,j-1} + \beta_j (x_{ij}+1) + \beta_{j+1} x_{i,j+1} + \dots + \beta_p x_{ip}\}$$

Hence if we study the ODDS RATIO

$$\frac{\left(\frac{\pi_k}{1-\pi_k}\right)}{\left(\frac{\pi_i}{1-\pi_i}\right)} = \frac{\exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{i,j-1} + \beta_j (x_{ij}+1) + \beta_{j+1} x_{i,j+1} + \dots + \beta_p x_{ip}\}}{\exp\{\beta_1 + \beta_2 x_{i2} + \dots + \beta_{j-1} x_{i,j-1} + \beta_j x_{ij} + \beta_{j+1} x_{i,j+1} + \dots + \beta_p x_{ip}\}}$$

$$= \exp\{\cancel{\beta_1} + \cancel{\beta_2 x_{i2}} + \dots + \cancel{\beta_{j-1} x_{i,j-1}} + \beta_j (x_{ij}+1) + \cancel{\beta_{j+1} x_{i,j+1}} + \dots + \cancel{\beta_p x_{ip}} - \cancel{\beta_1} - \cancel{\beta_2 x_{i2}} - \dots - \cancel{\beta_{j-1} x_{i,j-1}} - \beta_j x_{ij} - \cancel{\beta_{j+1} x_{i,j+1}} - \dots - \cancel{\beta_p x_{ip}}\}$$

$$= \exp\{\beta_j x_{ij} + \beta_j - \beta_j x_{ij}\} = \exp\{\beta_j\}$$

$$\Rightarrow \frac{\pi_k}{1-\pi_k} = e^{\beta_j} \frac{\pi_i}{1-\pi_i}$$

If we increase the covariate x_j by one unit, the ODDS CHANGE BY A MULTIPLICATIVE FACTOR e^{β_j} (keeping all other covariates fixed).

Moreover if we compute the log

$$\log \frac{\pi_k}{1-\pi_k} = \beta_j + \log \frac{\pi_i}{1-\pi_i} \Rightarrow \beta_j = \log \frac{\pi_k}{1-\pi_k} - \log \frac{\pi_i}{1-\pi_i}$$

The coefficient β_j represents the (additive) CHANGE IN THE LOG-ODDS if we increase the covariate x_j by 1 unit, keeping all other covariates fixed.

- INTERPRETATION WITH A BINARY COVARIATE

(2×2 contingency table)

consider a logistic regression with only one covariate, and that such covariate is binary.

e.g.: study about the efficacy of a treatment

$$y_i = \begin{cases} 1 & \text{alive} \\ 0 & \text{dead} \end{cases} \qquad z_i = \begin{cases} 1 & \text{treatment} \\ 0 & \text{placebo} \end{cases}$$

We can express the data in a 2×2 contingency table.
Each cell contains the counts of individuals with the corresponding combination of (y_i, z_i)

	$z_i = 1$	$z_i = 0$
$y_i = 1$	#(1,1)	#(1,0)
$y_i = 0$	#(0,1)	#(0,0)

model: $Y_i \sim \text{Bernoulli}(\pi_i)$ $\quad \pi_i = \frac{e^{\beta_1 + \beta_2 z_i}}{1+e^{\beta_1+\beta_2 z_i}}$

Consider an individual i that received the treatment

$(\pi_i \mid z_i = 1) = \mathbb{P}(Y_i = 1 \mid z_i = 1) = \frac{e^{\beta_1+\beta_2}}{1+e^{\beta_1+\beta_2}}$ → probability of surviving, having received the treatment

and $(1-\pi_i \mid z_i = 1) = \mathbb{P}(Y_i = 0 \mid z_i = 1) = \frac{1}{1+e^{\beta_1+\beta_2}}$ → probability of not surviving, having received the treatment

odds for an individual that received the treatment

$$\left(\frac{\pi_i}{1-\pi_i} \,\Big|\, z_i = 1\right) = e^{\beta_1+\beta_2}$$

Consider now that individual i received instead the placebo

$(\pi_i \mid z_i = 0) = \mathbb{P}(Y_i = 1 \mid z_i = 0) = \frac{e^{\beta_1}}{1+e^{\beta_1}}$ → probability of surviving, having received the placebo

and $(1-\pi_i \mid z_i = 0) = \mathbb{P}(Y_i = 0 \mid z_i = 0) = \frac{1}{1+e^{\beta_1}}$ → probability of not surviving, having received the placebo

odds for an individual that received the placebo

$$\left(\frac{\pi_i}{1-\pi_i} \,\Big|\, z_i = 0\right) = e^{\beta_1}$$

The ODDS RATIO is

$$\frac{\left(\frac{\pi_i}{1-\pi_i} \,\Big|\, z_i = 1\right)}{\left(\frac{\pi_i}{1-\pi_i} \,\Big|\, z_i = 0\right)} = \frac{\frac{\mathbb{P}(Y_i=1 \mid z_i=1)}{\mathbb{P}(Y_i=0 \mid z_i=1)}}{\frac{\mathbb{P}(Y_i=1 \mid z_i=0)}{\mathbb{P}(Y_i=0 \mid z_i=0)}} = e^{\beta_2}$$

The odds using a placebo are multiplied by a factor e^{β_2} to obtain the odds using the treatment.

Or equivalently

$$\log \left[\frac{\frac{\mathbb{P}(Y_i=1 \mid z_i=1)}{\mathbb{P}(Y_i=0 \mid z_i=1)}}{\frac{\mathbb{P}(Y_i=1 \mid z_i=0)}{\mathbb{P}(Y_i=0 \mid z_i=0)}}\right] = \beta_2$$

$$\Rightarrow \log \frac{\mathbb{P}(Y_i=1 \mid z_i=1)}{\mathbb{P}(Y_i=0 \mid z_i=1)} = \log \frac{\mathbb{P}(Y_i=1 \mid z_i=0)}{\mathbb{P}(Y_i=0 \mid z_i=0)} + \beta_2$$