MATH 60604A
Statistical modelling
§ 4b - Logistic regression

HEC Montréal
Department of Decision Sciences

Generalized linear model for binary variables

- In the case of a binary response variable, assume Y_i follows a Bernoulli distribution with parameter π_i, $Y_i \sim \text{Bin}(\pi_i)$, where

$$\pi_i = \mathsf{P}\left(Y_i = 1 \mid \mathbf{X}_i\right) = \mathsf{E}\left(Y_i \mid \mathbf{X}_i\right).$$

- An appropriate link function for binary responses is the logit function

$$g(z) \coloneqq \mathrm{logit}(z) = \ln\left(\frac{z}{1-z}\right).$$

- The **logistic regression model** is

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \coloneqq \beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}.$$

- The logit function g is the **quantile function of the logistic distribution** and links $\mathsf{E}\left(Y_i \mid \mathbf{X}_i\right) = \pi_i(\mathbf{X}_i)$ and η_i.

Logistic regression: logit function

- The logistic model is

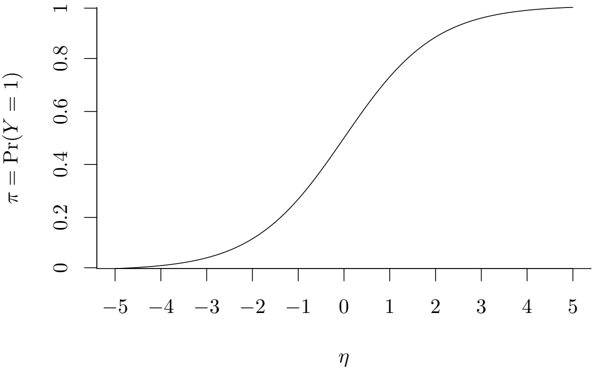
$$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip}.$$

- This model can also be written on the mean scale by using the inverse-logit (expit) function,

$$\mathsf{E}\left(Y_i \mid \mathbf{X}_i\right) = \pi_i = \frac{\exp(\beta_0 + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip})}{1 + \exp(\beta_0 + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip})}.$$

- We have an expression for the mean $\pi_i = \mathsf{E}\left(Y_i \mid \mathbf{X}_i\right)$ as a function of the explanatory variables $\mathbf{X}_i$, but...
  - what does this function look like?
  - what does this tell us about the relationship between π_i and η_i (and thus $\mathbf{X}_i$)?

Logistic distribution function



- Notice π is an **increasing function** of $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j \mathrm{X}_j$.
  - If β_j is positive and X_j increases, $\mathrm{P}\left(Y = 1\right)$ also increases.
  - If β_j is negative and X_j increases, $\mathrm{P}\left(Y = 1\right)$ decreases.
- We also see that the relationship between $\mathrm{P}\left(Y = 1\right)$ and η (and thus each X_j) is non-linear.

Parameter interpretations in terms of odds

- Quantifying the effect sizes in logistic regression is not easy because it's a nonlinear model.
- The coefficients can be interpreted in terms of odds and odds ratios.
- Let $\pi = \mathsf{P}\left(Y = 1 \mid \mathrm{X}_1, \ldots, \mathrm{X}_p\right)$, the logistic regression model is

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p.$$

- By exponentiating both sides, we obtain

$$\mathrm{odds}(Y \mid \mathbf{X}) = \frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})} = \exp(\beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p),$$

where $\pi(\mathbf{X})/\{1-\pi(\mathbf{X})\}$ are the odds of $\mathsf{P}\left(Y = 1 \mid \mathbf{X}\right)$ relative to $\mathsf{P}\left(Y = 0 \mid \mathbf{X}\right)$.

Odds

- The logit function corresponds to modelling the **log-odds.**
- The odds for binary Y are the quotient

$$\text{odds}(\pi) = \frac{\pi}{1-\pi} = \frac{\mathrm{P}\left(Y=1\right)}{\mathrm{P}\left(Y=0\right)}.$$

- For example, an odds of 4 means that the probability that $Y=1$ is four times higher than the probability that $Y=0$.
- An odds of 0.25 means the probability that $Y=1$ is only a quarter times the probability that $Y=0$, or equivalently, the probability that $Y=0$ is four times higher than the probability that $Y=1$.

$\mathrm{P}\left(Y=1\right)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Odds	0.11	0.25	0.43	0.67	1	1.5	2.33	4	9
Odds (frac.)	$\frac{1}{9}$	$\frac{1}{4}$	$\frac{3}{7}$	$\frac{2}{3}$	1	$\frac{3}{2}$	$\frac{7}{3}$	4	9

Interpretation of the intercept in terms of the odds

- When $X_1 = \cdots = X_p = 0$, it is clear that

$$\text{odds}(Y \mid \mathbf{X} = \mathbf{0}_p) = \exp(\beta_0)$$

  and

$$P\left(Y = 1 \mid X_1 = 0, \ldots X_p = 0\right) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

  which represents the probability that $Y = 1$ when $\mathbf{X} = \mathbf{0}_p$.
- As for linear regression, $X_1 = \cdots = X_p = 0$ might not be physically possible, in which case there is no sensible interpretation for β_0.

Parameter interpretation in terms of the odds ratio

Consider for simplicity a logistic model of the form

$$\operatorname{logit}(\pi) = \beta_0 + \beta_1 x.$$

The factor $\exp(\beta_1)$ is the change in odds when X increases by one unit,

$$\text{odds}(Y \mid \mathrm{X} = x + 1) = \exp(\beta_1) \times \text{odds}(Y \mid \mathrm{X} = x).$$

- If $\beta_1 = 0$ then the odds ratio is unity,
  meaning that the variable X is not associated with the odds of Y
- If β_1 is positive, then the odds ratio $\exp(\beta_1)$ is larger than one,
  meaning that, as X increases, the odds of Y increases.
- If β_1 is negative, the odds ratio $\exp(\beta_1)$ is smaller than one,
  meaning that, as X increases, the odds of Y decreases.

Note that, when there are several explanatory variables in the model, the interpretation of β_1 is when all other variables in the model are held constant.

Interpretation of β_k in terms of odds ratio

For the logistic model, the odds ratio when $X_k = x_k + 1$ versus $X_k = x_k$ when $X_j = x_j$ ($j = 1, \dots, p, j \neq k$) is

$$\frac{\text{odds}(Y \mid X_k = x_k + 1, X_j = x_j, j \neq k)}{\text{odds}(Y \mid X_k = x_k, X_j = x_j, j \neq k)} = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j + \beta_k\right)}{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)}$$
$$= \exp(\beta_k).$$

When X_k increases by one unit **and all the other covariates are held constant**, the odds of Y changes by a factor $\exp(\beta_k)$.

- The odds increase if $\exp(\beta_k) > 1$, i.e., if $\beta_k > 0$.
- The odds decrease if $\exp(\beta_k) < 1$, i.e., if $\beta_k < 0$.

The effect of β_k is larger when π is near 0.5 than near endpoints of $(0, 1)$.