Logistic Regression

Recall: linear regression is used when response variable is quantitative, and ideally when error distribution is normal and linearity is a good assumption. However, in practice, other types of response arise. Two examples:

(a) response is binary/categorical/qualitative (e.g. presence/absence of a disease)

(b) response occurs as counts/integer (e.g. arrivals in a queue)

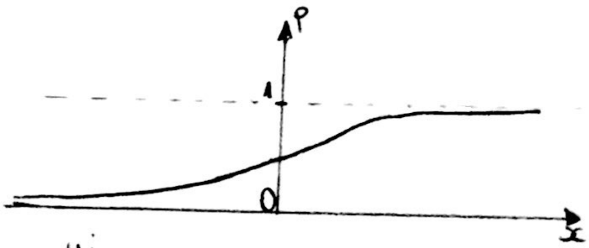
Today we're going to study models of type (a), which can be modeled by Logistic Regression, and response is binary / binomial

- Simple Logistic Regression $Y \in \{0, 1\}$ <=> binary / Bernoulli
- Multinomial Logistic Regression $Y \in \{0, 1, 2, \ \}$ <=> Binomial

Why Logistic Regression?

(1) When you must have $0 \leq E[Y|X] \leq 1$

(2) s-shaped curve arises empirically



logit (logistic) function $= \overbrace{\ln\left(\frac{p}{1-p}\right)}^{y} = \beta_0 + \beta_1 x$ or $\overbrace{\ln\left(\frac{p_i}{1-p_i}\right)}^{y_i} = \beta_0 + \beta_1 x_i$ (one covariate)

or $\overbrace{\ln\left(\frac{p_i}{1-p_i}\right)}^{y_i} = \beta_0 + \sum_j \beta_j x_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}$ (j covariates)

- * log odds = logit function $= \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$
- * odds $= \frac{p}{1-p} = e^{\beta_0 + \beta_1 x} = e^{\text{logit function}}$
- * odds ratio (A vs B) $= \frac{e^{\beta_0 + \beta_1 A}}{e^{\beta_0 + \beta_1 B}} = e^{\beta_1 (A - B)}$

Interpretation: for any value of x, increasing the covariate x by one unit increases the log odds by β_1.

$$\Rightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 x} \Rightarrow p = e^{\beta_0 + \beta_1 x} - p\,e^{\beta_0 + \beta_1 x}$$

$$\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

where the last equality comes from: $\frac{\cancel{e^{\beta_0 + \beta_1 x}}}{\cancel{e^{\beta_0 + \beta_1 x}}} \frac{(1)}{(1/e^{\beta_0 + \beta_1 x} + 1)} = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$

Model Estimation using MLE - Response is Binary $\therefore Y_i \in \{0,1\}$
(Simple Logistic Regression)

① For p ($\hat{p}_{MLE}$) $\quad Y_i | x_{1i}, x_{ki} \sim \text{Bernoulli}(p)$

Likelihood: $L(y_i; p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$

loglikelihood: $\ell(y_i; p) = \sum_{i=1}^{n} \{ y_i \log p + (1-y_i)\log(1-p) \}$

$\underset{p}{\text{argmax}}\ \ell(y_i; p) = ?$ $\Rightarrow$ Take 1st derivatives wrt p and set to 0

$$\frac{\partial \ell(y_i; p)}{\partial p} = \sum_{i=1}^{n} \frac{y_i}{p} - \frac{1-y_i}{1-p} = 0 \Leftrightarrow \frac{\sum y_i}{p} = \frac{n - \sum y_i}{1-p}$$

$$\Leftrightarrow \sum y_i - p\sum y_i = pn - p\sum y_i$$

$$\boxed{\hat{p}_{MLE} = \frac{\sum y_i}{n} = \bar{y}}$$

② for β ($\hat{\beta}_{MLE}$ $\Rightarrow$ no closed form $\Rightarrow$ solve numerically)

loglikelihood
$$\ell(y_i; \beta_0, \beta) = \sum_{i=1}^{n} \left\{ y_i \log \underbrace{\left(\frac{1}{1+e^{-(\beta_0 + x_i^T\beta)}} \right)}_{h(x_i;\beta_0,\beta)\ = p} + (1-y_i) \log \underbrace{\left(1 - \frac{1}{1+e^{-(\beta_0 + x_i^T\beta)}} \right)}_{1-h(x_i;\beta_0,\beta)\ = 1-p} \right\}$$

notice: $h(x_i; \beta_0, \beta) = \dfrac{1}{1+e^{-(\beta_0 + x_i^T\beta)}}$

$$\Rightarrow \log h(x_i; \beta_0, \beta) = \log 1 - \log\left(1 + e^{-(\beta_0 + x_i^T\beta)}\right) = -\log\left(1+e^{-(\beta_0 + x_i^T\beta)}\right)$$

$\rightarrow$ loglikelihood
$$\ell(y_i; \beta_0, \beta) = \sum_{i=1}^{n} \left\{ -y_i \log\left(1+e^{-(\beta_0 + x_i^T\beta)}\right) + (1-y_i)\log\left(\frac{e^{-(\beta_0 + x_i^T\beta)}}{1+e^{-(\beta_0 + x_i^T\beta)}} \right) \right\}$$

Now, to apply Newton's Method (to get approx. $\hat{\beta}$) we need to derive gradients and Hessian.

* Getting Gradients :

chain rule $(\log x)' \circ x'$

$$\nabla \ell(\beta_0) = \sum_{i=1}^{n} \left\{ y_i \frac{e^{-(\beta_0 + x_i^T\beta)}}{1+e^{-(\beta_0 + x_i^T\beta)}} + (1-y_i) \frac{1}{\cancel{\frac{e^{-(\beta_0+x_i^T\beta)}}{1+e^{-(\beta_0+x_i^T\beta)}}}} \frac{\left(-e^{-(\beta_0+x_i^T\beta)}\right)\left(1+e^{-(\beta_0+x_i^T\beta)}\right) - \left(-e^{-(\beta_0+x_i^T\beta)}\, e^{-(\beta_0+x_i^T\beta)}\right)}{\left(1+e^{-(\beta_0+x_i^T\beta)}\right)^{\cancel{2}}} \right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i \frac{e^{-(\beta_0+x_i^T\beta)}}{1+e^{-(\beta_0+x_i^T\beta)}} + (1-y_i) \frac{1}{\cancel{e^{-(\beta_0+x_i^T\beta)}}} \left(-\cancel{e^{-(\beta_0+x_i^T\beta)}} + \frac{\left(e^{-(\beta_0+x_i^T\beta)}\right)^{\cancel{2}}}{1+e^{-(\beta_0+x_i^T\beta)}} \right) \right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i \frac{e^{-(\beta_0+x_i^T\beta)}}{1+e^{-(\beta_0+x_i^T\beta)}} + (1-y_i) \left(-1 + \frac{e^{-(\beta_0+x_i^T\beta)}}{1+e^{-(\beta_0+x_i^T\beta)}} \right) \right\}$$

$$= \sum_{i=1}^{n} \left\{ -1 + \frac{e^{-(\beta_0+x_i^T\beta)}}{1+e^{-(\beta_0+x_i^T\beta)}} + y_i \right\} = \sum_{i=1}^{n} \left\{ y_i - \frac{1}{(1+e^{-(\beta_0+x_i^T\beta)})} \right\} = \sum_{i=1}^{n} \left\{ y_i - h(x_i, \beta_0, \beta) \right\} //$$

$$\nabla \ell(\beta) = \sum_{i=1}^{n} \left\{ y_i \frac{e^{-(\beta_0+x_i^T\beta)} x_i}{1+e^{-(\beta_0+x_i^T\beta)}} + (1-y_i) \frac{1}{\cancel{\frac{e^{-(\beta_0+x_i^T\beta)}}{1+e^{-(\beta_0+x_i^T\beta)}}}} \frac{\left(-\cancel{e^{-(\beta_0+x_i^T\beta)}}\right)\left(1+e^{-(\beta_0+x_i^T\beta)}\right) - \left(-\cancel{e^{-(\beta_0+x_i^T\beta)}}\, e^{-(\beta_0+x_i^T\beta)}\right)}{\left(1+e^{-(\beta_0+x_i^T\beta)}\right)^{\cancel{2}}} x_i \right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i \frac{e^{-(\beta_0+x_i^T\beta)} x_i}{1+e^{-(\beta_0+x_i^T\beta)}} + (1-y_i) \left(-x_i + \frac{e^{-(\beta_0+x_i^T\beta)} x_i}{1+e^{-(\beta_0+x_i^T\beta)}} \right) \right\}$$

$$= \sum_{i=1}^{n} \left\{ -x_i + \frac{e^{-(\beta_0+x_i^T\beta)} x_i}{1+e^{-(\beta_0+x_i^T\beta)}} + y_i x_i \right\} = \sum_{i=1}^{n} \left\{ x_i \left(\frac{e^{-(\beta_0+x_i^T\beta)}}{1+e^{-(\beta_0+x_i^T\beta)}} - 1 + y_i \right) \right\}$$

$$= \sum_{i=1}^{n} \left\{ x_i \left(-\frac{1}{1+e^{-(\beta_0+x_i^T\beta)}} + y_i \right) \right\} = \sum_{i=1}^{n} \left\{ x_i \left(y_i - h(x_i; \beta_0, \beta) \right) \right\} //$$

$$\boxed{\text{Hence, } \nabla \ell(\beta_0, \beta) = \sum_{i=1}^{n} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \left(y_i - h(x_i; \beta_0, \beta) \right)}$$

where $h(x_i; \beta_0, \beta) = \dfrac{1}{1+e^{-(\beta_0+x_i^T\beta)}} = \dfrac{e^{\beta_0+x_i^T\beta}}{1+e^{\beta_0+x_i^T\beta}} = p$

* Getting Hessian Matrix

$$\nabla^2 l(\beta_0) = \frac{\partial^2 l(\beta_0,\beta)}{\partial \beta_0^2} = \sum_{i=1}^{n} \left\{ -\left(\frac{e^{-(\beta_0 + x_i^T\beta)}}{(1+e^{-(\beta_0+x_i^T\beta)})^2} \right) \right\} = \sum_{i=1}^{n} \left\{ \overbrace{\frac{1}{1+e^{-(\beta_0+x_i^T\beta)}}}^{h} \overbrace{\left(\frac{-e^{-(\beta_0+x_i^T\beta)}}{1+e^{-(\beta_0+x_i^T\beta)}} \right)}^{(h-1)} \right\}$$

$$= \sum_{i=1}^{n} \left\{ h(x_i;\beta,\beta_0)\left(h(x_i;\beta,\beta_0) - 1 \right) \right\} //$$

$$\nabla^2 l(\beta) = \frac{\partial^2 l(\beta_0,\beta)}{\partial \beta \partial \beta^T} = \sum_{i=1}^{n} \left\{ -x_i \left(\frac{x_i^T (e^{-(\beta_0+x_i^T\beta)})}{(1+e^{-(\beta_0+x_i^T\beta)})^2} \right) \right\} = \sum_{i=1}^{n} \left\{ x_i x_i^T h(x_i;\beta,\beta_0)\left(h(x_i;\beta_0,\beta) - 1 \right) \right\}$$

$$\nabla^2 l(\beta_0,\beta) = \frac{\partial^2 l(\beta_0,\beta)}{\partial \beta_0 \partial \beta^T} = \sum_{i=1}^{n} \left\{ -\left(\frac{x_i^T e^{-(\beta_0+x_i^T\beta)}}{(1+e^{-(\beta_0+x_i^T\beta)})^2} \right) \right\} = \sum_{i=1}^{n} \left\{ x_i^T h(x_i;\beta_0,\beta)\left(h(x_i;\beta_0,\beta) - 1 \right) \right\}$$

and $\frac{\partial^2 l(\beta_0,\beta)}{\partial \beta \partial \beta_0} = \sum_{i=1}^{n} \left\{ x_i h(x_i;\beta_0,\beta)\left(h(x_i;\beta_0,\beta) - 1 \right) \right\}$

Hence, the Hessian:

$$\nabla^2(\beta_0,\beta) = \sum_{i=1}^{n} h(x_i;\beta_0,\beta)\left(h(x_i;\beta_0,\beta) - 1 \right) \begin{bmatrix} 1_{1\times 1} & x^T_{i_{(1\times K)}} \\ x_{i_{(K\times 1)}} & x_i x^T_{i_{(K\times K)}} \end{bmatrix}_{(K+1)(K+1)}$$

or equivalently $\nabla^2(\beta_0,\beta) = \sum_{i=1}^{n} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \begin{bmatrix} 1 \\ x_i^T \end{bmatrix}^T h(x_i;\beta_0,\beta)\left(h(x_i;\beta_0,\beta) - 1 \right)$

* Applying Newton's Method

$$\hat{\beta}^{NEW} = \hat{\beta}^{OLD} - [\text{Hessian}]^{-1} \cdot \text{gradient}$$

$$\begin{bmatrix} \beta_0 \\ \beta \end{bmatrix}^{(t+1)} = \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix}^{(t)} - \underbrace{\mu}_{\text{step}} \left[\nabla^2 \ell(\beta_0^{(t)}, \beta^{(t)}) \right]^{-1} \cdot \nabla \ell(\beta_0^{(t)}, \beta^{(t)})$$

define $\begin{bmatrix} \beta_0 \\ \beta \end{bmatrix}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and choose a step (e.g. 0.01, 1, etc) to get convergence

→ Another common notation:

if we write the loglikelihood as:

$$\ell(y;\beta) = -Y^T \log(1 + e^{-X^T\beta}) + (1^T - Y^T) \log\left(\frac{e^{-X^T\beta}}{1+e^{-X^T\beta}}\right)$$

Then $\nabla \ell(y;\beta) = X^T(Y - P)$ where P is the vector $\begin{pmatrix} h(x_1;\beta) \\ h(x_2;\beta) \\ \vdots \\ h(x_n;\beta) \end{pmatrix}$

And $\nabla^2 \ell(y;\beta) = -X^T W X$ where W is the matrix of weights: $W = [P(x_i)(1-P(x_i))]_{n \times n}$

Therefore, we rewrite Newton's Method as:

$$\boxed{\hat{\beta}^{new} = \hat{\beta}^{OLD} + (X^T W X)^{-1} X^T (Y - P)}$$

* Standard error $se(\hat{\beta}) = (X^T W X)^{-1}$