[quick background on Logit]

Logit Model

Consider a dichotomous ~~and~~ response variables Y with two measurement levels.

Let $\pi(x) = P(Y=1 \mid X=x) = (1-P(Y=0 \mid X=x)$

odds ratio $\dfrac{\pi(x)}{1-\pi(x)}$

$$\log(\text{odds ratio}) = \text{Logit}\,[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$
$$= \alpha + \beta x$$

Thus the odds $= \exp(\alpha + \beta x)$

- Rate of increase or decrease of the S shaped curve of $\pi(x)$
- The sign of 'β' indicates whether curve ascends $(\beta > 0)$ or descends $(\beta < 0)$

Multiple Logit Model.

Let 'k' denote nos of predictors for a binary response 'Y' by $x_1, x_2 \dots x_k$.

Then we have. $\text{Logit}\,[P(Y=1)] = \alpha + \beta_1 x_1 + \dots \beta_k x_k$.

or $$\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + \dots \beta_k x_k.$$

$$\pi(x) = \frac{\exp(\alpha + \beta_1 x_1 + \dots \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \dots \beta_k x_k)}$$

Multinomial Logit

Consider the following.

- 'n' independent observations
- 'p' explanatory variables.
- 'k' Categories.

Take any category as base. → Let us take category 'J' as base.

- π_{ij} → multinomial probability of 'ith' observation falling in the 'j'th category

Then $$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + x_i' \beta_j$$

We assume that the log of odds (wrt to base category) follows a linear model.

Linear predictor:

Consider the linear predictor function $f(k,i)$ to predict the probability that observation 'i' has outcome 'k'

$$f(k,i) = \beta_{0,k} + \beta_{1,k} x_{1,i} + \beta_{2,k} x_{2,i} + \cdots \beta_{M,k} x_{M,i}$$

$\beta_{m,k}$ → Regression coefficient for mth explanatory variable and 'k'th outcome

Writing more compactly we have.

$$\cancel{f(k,i)} \quad f(k,i) = \beta_k \cdot X_i$$

As a set of independent binary regressions:

For 'K' possible outcomes, run K-1 independent binary Logistics regression models.

one outcome 'say the last' is chosen as pivot..

Thus we have 'K-1' equations as follows.

$$\ln \frac{Pr(Y_i=1)}{Pr(Y_i=K)} = \beta_1 \cdot X_i \Rightarrow Pr(Y_i=1) = Pr(Y_i=K)e^{\beta_1 \cdot X_i}$$

$$\vdots$$

$$\ln \frac{Pr(Y_i=K-1)}{Pr(Y_i=K)} = \beta_{K-1} \cdot X_i$$

Sum of probabilities must = 1 : Thus we have.

$$Pr(Y_i=K) = 1 - \sum_{k=1}^{K-1} Pr(Y_i=k)$$

$$= 1 - \sum_{k=1}^{K-1} Pr(Y_i=K)\, e^{\beta_k \cdot X_i}$$

$$\Rightarrow Pr(Y_i=K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}} \quad \text{— (A)}$$

We can Now use (A) to find other Probabilities.

$$Pr(Y_i=1) = \frac{e^{\beta_1 \cdot X_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot X_i}}$$

Background : Recap Logistics Regression

π_i = Probability of success for any given obs.

odds ratio = $\dfrac{\pi_i}{1-\pi_i}$

$$\text{Logit} = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=0}^{K} x_{ik}\beta_k \quad , i=1,2,\cdots N.$$

Solving for $\pi_j = P(x)$

we have $$\frac{P(x)}{1-P(x)} = \exp\left(\sum_{j=0}^{k} x_j \cdot \beta_j\right) = \exp(z) \text{ say.}$$

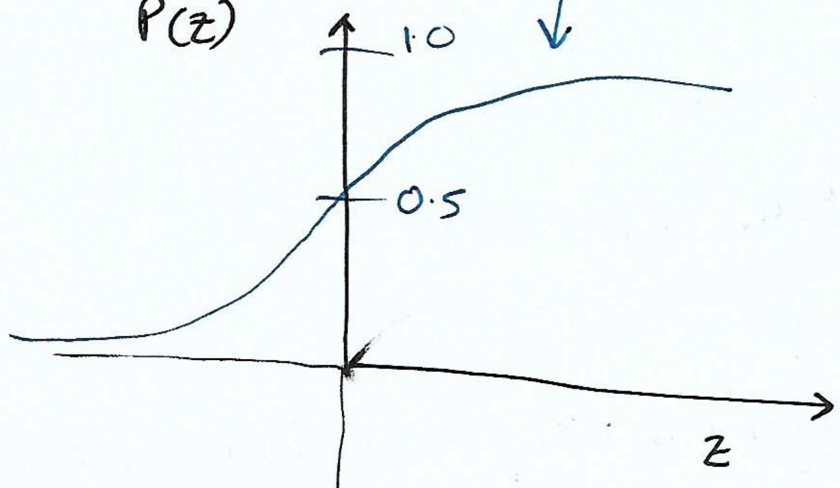
$$\therefore \frac{P(x)}{1-P(x)} = \exp(z)$$

$$\therefore P(x) = \frac{\exp(z)}{1+\exp(z)}$$

or $$P(z) = \frac{\exp(z)}{1+\exp(z)} = \frac{1}{1+\exp(-z)}$$

Sigmoid Function.

Maps a real line to (0,1)



A-std. Proof: $P'(z) = P(z)(1-P(z))$

we have $P(z) = \frac{1}{1-e^{-z}} = (1-e^{-z})^{-1}$

$\therefore\ P'(z) = -1(1-e^{-z})^{-2}(-1)(-e^{-z})$

$\therefore\ P'(z) = \frac{-e^{-z}}{(1-e^{-z})(1-e^{-z})}$

Now $P(z) = \frac{1}{(1-e^{-z})}$ and $1 - P(z)$

$= \frac{1 - \frac{1}{1-e^{-z}}}{1} = \frac{(1-e^{-z}) - 1}{(1-e^{-z})} = \frac{-e^{-z}}{(1-e^{-z})}$

$\therefore\ \left[P'(z) = P(z)(1-P(z))\right]$

Determining the coefficients:] Maximizing the Log Likelihood.

Likelihood Function for Logistics Regression. →

- For each training data point. we have a vector of features x_i and an observed class y_i

Prob of that class = p if $y_i = 1$, or $1-p$ if $y_i = 0$

The likelihood Function is →

: Product of predicted probabilities of the N individual observations

The likelihood is written as

$$L(\beta_0,\beta) = \prod p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$$

The log likelihood is written as

$$lL(\beta_0,\beta) = \sum_{i=1}^{N}\left[y_i \log p(x_i) + (1-y_i)\log(1-p(x_i))\right]$$

Maximize the log likelihood.

We take derivative wrt β'

$$\nabla_b(\ell L) = \sum_{i=1}^{N} y_i \frac{P_i'}{P_i} x_i + \sum_{i=1}^{N} (1-y_i) \frac{P_i' x_i}{1-P_i}$$

Now $P_i' = P_i(1-P_i)$

$$\therefore \nabla_b(\ell L) = \sum_{i=1}^{N} \left[y_i(1-P_i) - (1-y_i)P_i \right] x_i$$

$\Rightarrow$ Setting $\sum_{i=1}^{N} y_i x_i - P_i x_i = 0$

How do we solve for coefficients.

Say we have a vector valued function. $y = f(b)$.
We want $f(b_{opt}) = 0$. Assume we start with initial guess. (b_0)

$$f(b_0 + \Delta) \simeq f(b_0) + \Delta f'(b_0) = 0$$

$$\therefore \Delta_0 = -\frac{f(b_0)}{f'(b_0)}$$

upgrade rule

$$b_1 = b_0 + \Delta_0$$

Recap:
Linear Regression

$y = x^T b$

$xy = xx^T b$

$b = (xx^T)^{-1} xy$

Now we have $f = \nabla_b(\ell L) = \sum y_i x_i - P_i x_i = 0$

also $H = \frac{\partial}{\partial b}(\nabla_b \ell\ell) = -\sum x_i(\nabla_b(P_i))$

$$= -\sum x_i P_i(1-P_i) x_i^T$$

$$= XWX^T$$

Now in matrix form.

$$\nabla b(\ell L) = X(y - P_k)$$

or $\left[\Delta_k = (XW_kX^T)^{-1} X (y - P_k) \right]$

Log likelihood For Multinomial case:

Likelihood $L = \prod_{i=1}^{n} \prod_{h=0}^{r} p_{ih}^{y_{ih}}$

$y_{ih} \to$ observed values

$p_{ih} \to$ theoretical values.

$\therefore$ Log likelihood $LL = \sum_{i=1}^{n} \sum_{h=0}^{r} y_{ih} \ln p_{ih}$ — (A)

Let $B_h = \cancel{b_{hj}}\ b_{hj}$ be the $(k+1) \times 1$ col vector of binary logistics reg. coefficients of the outcome 'h' compared to '0'

Let B be the $r(k+1) \times 1$ col. vector consisting of $B_0 \ldots B_r$ arranged in a column.

Let X be the design matrix $n \times (k+1)$

For outcomes 'h' and 'l' let V_{hl} be the $n \times n$ diag matrix whose main diag. contains elements of the form.

$$V_{ii} = \begin{cases} p_{ih}(1-p_{ih}) & h = l \\ -p_{ih}p_{il} & h \neq l \end{cases}$$

Let $C_{hl} = X^T V_{hl} X$. Now define the $n \times$ $nr \times nr$ matrices.

$$C = \begin{bmatrix} C_{11} & \cdots & C_{1r} \\ C_{r1} & & C_{rr} \end{bmatrix}$$

Then $S = C^{-1}$ is the Covariance matrix for B.

For max (A) log likelihood.

We have. $\sum_{i=1}^{n}(y_{ih} - p_{ih}) = 0$ and $\sum_{i=1}^{n} x_{ij}\ (y_{ih} - p_{ih}) = 0$

Thus we get the following matrix eq.

$$X^T (y - p) = 0$$

Let B^0 be initial guess for B

For each mth iter. we have.

$$B^{m+1} = B^{(m)} + S^{-1} X^T (y - p(m))$$

For sufficiently Large m.

$B^{(m+1)} \simeq B(m)$ is a good approx. For 'B'