GLMs (Generalized Linear Models) Introduction

1. Assumption

Distribution of the dependent variable (for example $Y \sim Bin(n, p)$, $Y \sim Gamma(\alpha, \beta)$)

2. Specify a link function $g(\cdot)$

("Linearize Y") such that $g(E(Y)) = X\beta$

If $E[y_i] = \mu_i$, $g(\mu_i) = \eta_i$, $\eta_i = X_i\beta$

Fitting a GLM

Suppose $Y_i \sim Bin(n_i, p_i)$.

and we wish to predict Y_i / n_i.

Then $E(Y_i / n_i) = p_i$

A common link function is logit: $\log\left(\frac{p}{1-p}\right)$

$$g(\mu_i) = \log\left(\frac{\mu_i}{n_i - \mu_i}\right), \quad \mu_i = n_i p_i$$

$$P(Y_i = y_i | p_i, n_i) = \binom{n_i}{y_i} p_i^{y_i} (1-p)^{n_i - y_i}$$

Suppose now that you know $y_i \overset{iid}{\sim} Bin(n, p)$

'What is the likelihood function?

$$L = \prod_{i=1}^{K} \binom{n}{y_i} \cdot p^{y_i} (1-p)^{n - y_i}$$

$$\log L = \sum_{i=1}^{K} y_i \log p + (n - y_i) \log(1-p) + \log \binom{n}{y_i}$$

What value of p maximizes the likelihood?

$$\hat{p} = \underset{p \in [0,1]}{\arg\max} \log L$$

Great! How do we connect this with $X\beta$ —?

Remmember the logit function

$$X_i \beta = \log\left(\frac{\mu_i}{n - \mu_i}\right) = \log\left(\frac{np}{n - np}\right)$$

$$\Rightarrow p = \text{inv logit}(X_i\beta) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$$

$$\Rightarrow \underbrace{p = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}}$$

$\hookrightarrow$ substitute this into the log-Likelihood

$$\log L = \sum_{i=1}^{K}\left[\log\binom{n}{y_i} + y_i \cdot \log\left[\frac{1}{n}\cdot\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right] + (1-y_i)\cdot\log\left[1-\frac{1}{n}\cdot\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right]\right]$$

Now you can solve

$$\hat{\beta} = \underset{\beta}{\arg\max}\ \log L$$

The same logic is applied to the deviance

$$D = -2\left(\ell(\beta_{PM}) - \ell(\beta_{SM})\right)$$

D - deviance

$\ell(\beta_{PM})$: log-likelihood of the Proposed model (PM)

$\ell(\beta_{SM})$ log-likelihood of the saturated model (SM)

(one parameter per observation - fits the data perfectly) [n parameters to estimate]

$\ell(\beta_{SM}) = \underline{1}$

Residual deviance : $-2\ell(\beta_{PM})$

Null deviance : $-2\ell(\beta_{NM})$

NM - no model (empty model)

Exponential family of distributions

(normal, Binomial, poisson, gamma, exponential ...)

Density:

$$f(y;\theta,\phi) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y,\phi) \right\}$$

ϕ - dispersion parameter

θ - cannonical parameter

It can be shown that

$$E[Y] = b'(\theta) = \mu$$

(' denotes a derivative)

and $\mathrm{Var}[Y] = \phi\, b''(\theta) = \phi\, V(\mu)$

Link function

$$g(\mu_i) = X_i \beta$$

$$g = (b')^{-1} \Rightarrow g(\mu_i) = \theta_i$$

cannonical link fct.

The log-likelihood:

$$l = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i) \right]$$

The max. likelihood estimates are obtained by solving score equations

$$s(\beta_j) = \frac{\partial l}{\partial \beta_j} = 0 \quad , \text{for } \beta_j$$