GENERALISED LINEAR MODEL

EXPONENTIAL FAMILY

we have that if a distribution $f(y;\theta,\phi)$ is in an exponential family, we can write its PDF in the form:

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

where θ = canonical parameter
ϕ = dispersion parameter

A) • $y \sim N(\mu, \sigma^2)$, known σ^2

$$p(y) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \cdot \exp\left\{ -\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2} \right\}$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma} \cdot \exp\left\{ -\frac{1}{2\sigma^2} (y^2 - 2y\mu + \mu^2) \right\}$$

$$= \exp\left\{ \underbrace{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2}}_{\frac{y\theta - b(\theta)}{a(\phi)}} \underbrace{- \frac{1}{2\sigma^2} y^2 + \log\left(\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma}\right)}_{c(y,\phi)} \right\}$$

$\Rightarrow$ • canonical parameter $= \mu$
• $a(\phi) = \sigma^2$
• $b(\theta) = \frac{1}{2}\mu^2$

• $y = \frac{z}{N}$; $z \sim \text{binom}(N, p)$ for known N.

$$p(z) = \frac{n!}{(n-z)!\,z!} p^z (1-p)^{n-z}$$

also we have $\frac{\partial y}{\partial z} = \frac{1}{n}$, so:

$$p(y) = \frac{n!}{(n-yn)!\,(yn)!} \cdot p^{yn} \cdot (1-p)^{n-yn} \cdot \frac{1}{n}$$

$$= \exp\left\{ \log\binom{n}{ny} + ny \log p + N \log(1-p) - Ny \log(1-p) - \log N \right\}$$

$$= \exp\left\{ yn \log\left(\frac{p}{1-p}\right) - N \log\left(\frac{1}{1-p}\right) + \log\binom{n}{ny} - \log n \right\}$$

$\Rightarrow$ • $\theta = n \log\left(\frac{p}{1-p}\right)$

• $b(\theta) = N \log\left(\frac{1}{1-p}\right)$

• $a(\phi) = 1$

• $c(\phi) = \log\binom{n}{ny} - \log n$

• $y \sim \text{Poisson}(\lambda)$

$$p(y) = \frac{e^{-\lambda} \lambda^y}{\lambda}$$

$$= \exp\{ y \cdot \log(\lambda) - \lambda + (-\log(y!)) \}$$

$\Rightarrow$ • $\theta = \log(\lambda)$ $\quad b(\theta) = \lambda$

• $a(\phi) = 1$ $\quad c(\phi, y) = -\log(y!)$

B) We want to characterize the mean and variance of a distribution that belongs to the exponential family.

We define:

- $S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta)$

- $L(\theta) = \sum f(y_i, \theta)$

and we want to prove the score equations:

1) $E[S(\theta)] = 0$
2) $I(\theta) = VAR(S(\theta)) = -E(H(\theta))$

$\Downarrow$

1) $\int f(y,\theta)\, dy = 1$

$$\frac{\partial}{\partial \theta} \int f(y,\theta)\, dy = 0$$

$$\int \frac{\partial}{\partial \theta} f(y,\theta)\, dy = 0$$

$$\int \frac{\frac{\partial}{\partial \theta} f(y,\theta)}{f(y,\theta)} \cdot f(y,\theta)\, dy = 0$$

$$\int \underbrace{\frac{\partial}{\partial \theta} \log f(y,\theta)}_{S(\theta)} \cdot f(y,\theta)\, dy = E(S(\theta)) = 0$$

2) Start from previous result:

$$\int S(\theta) \cdot f(y,\theta) \cdot dy = 0$$

and derive again:

$$\int \frac{\partial}{\partial\theta} S(\theta) f(y,\theta) \cdot dy = 0$$

$$\int \frac{\partial}{\partial\theta} S(\theta) \cdot f(y,\theta)\, dy + S(\theta) \cdot \frac{\partial}{\partial\theta} f(y,\theta)\, dy = 0$$

$$\int H(\theta) \cdot f(y,\theta)\, dy = -\int S(\theta) \cdot f'(y,\theta)\, dy$$

$$E(H(\theta)) = -\int \frac{\partial}{\partial\theta} \log(f(y,\theta)) \cdot \frac{f'(y,\theta)}{f(y,\theta)} \cdot f(y,\theta)\, dy$$

$$= -\int \frac{\partial}{\partial\theta} \log f(y,\theta) \cdot \frac{\partial}{\partial\theta} \log f(y,\theta)\, dy$$

$$= -\int \left(\frac{\partial}{\partial\theta} \log f(y,\theta)\right)^2 dy$$

$$= -E(S(\theta)^2) \Rightarrow E(S(\theta)^2) = -E(H(\theta))$$

$\to 0$ from prev part

$$\text{now note: } V(S(\theta)) = \underbrace{E(S(\theta)^2)}_{-E(H(\theta))} - \cancel{E(S(\theta))}$$

c) We have to use the score equations to show that:

if $y \sim f(y|\theta, \phi)$ $\Rightarrow$ 1) $E(y) = b'(\theta)$

$\in$ exp family 2) $V(y) = a(\phi) \cdot b''(\theta)$

1) Recall $E(S(\theta)) = 0$

We have: $S(\theta) = \frac{\partial}{\partial \theta} \log f(y, \theta)$

$$= \frac{\partial}{\partial \theta}\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

$$= \frac{y}{a(\phi)} - \frac{b'(\theta)}{a(\theta)}$$

$$\Rightarrow E(S(\theta)) = 0 = E\left(\frac{y}{a(\phi)} - \frac{b'(\theta)}{a(\theta)}\right)$$

$$\rightarrow E(y) = \frac{b'(\theta)}{a(\phi)} \cdot a(\phi) = b'(\theta) \checkmark$$

2) Recall $VAR(S(\theta)) = -E(H(\theta))$ (*)

where: $VAR(S(\theta)) = V\left(\frac{y}{a(\phi)} - \frac{b'(\theta)}{a(\theta)}\right) = \frac{1}{a^2(\phi)} \cdot V(y)$

and

$$-E(H(\theta)) = -E\left(\frac{\partial}{\partial \theta}\left(\frac{y}{a(\phi)} - \frac{b'(\theta)}{a(\phi)}\right)\right) = -E\left(\frac{b''(\theta)}{a(\phi)}\right)$$

So that combining them through (*):

$$\frac{1}{a(\phi)^2} V(y) = \frac{b''(\theta)}{a(\phi)} \rightarrow V(y) = b''(\theta) \cdot a(\phi) \checkmark$$

D) The result can be easily used to compute the mean and the variance of $N(\mu, \delta^2)$

Indeed, by exploiting the result we found in A),
we have:

- $E(y) = b'(\theta) = \mu$
- $V(y) = b''(\theta) \cdot a(\phi) = 1 \cdot \delta^2 = \delta^2$

GENERALISED LINEAR MODELS

- PDF can be written as:

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i;\phi/w_i)\right\}$$

- for some invertible g function we have: $g(\mu_i) = x_i'\beta = \eta_i$

$$\Rightarrow E(y_i) = \mu_i$$

A) want to show that for a GLM we have:

☐ $\theta_i = (b')^{-1}(g^{-1}(x_i'\beta))$

we have:

$$g(\mu_i) = x_i'\beta \rightarrow \mu_i = g^{-1}(x_i'\beta) \rightarrow E(y_i) = g^{-1}(x_i'\beta)$$

however from exponential family we know:

$$E(y_i) = b'(\theta_i)$$

$$\Rightarrow b'(\theta_i) = g^{-1}(x_i'\beta) \longrightarrow \theta_i = (b')^{-1}(g^{-1}(x_i'\beta))$$

☐ $VAR(y_i) = \frac{\phi}{w_i} V(\mu_i)$ some function specified in terms of the building blocks of the exponential family

- from exponential family & GLM we have:

$$VAR(y_i) = b''(\theta_i)\cdot a(\phi) = \underbrace{b''(\theta_i)} \frac{\phi}{w_i}$$

$$= \frac{\phi}{w_i} \cdot b''\left(\overbrace{(b'^{-1})(g^{-1}(x_i'\beta))}\right)$$

$$= \frac{\phi}{w_i} \cdot b''((b'^{-1})(\mu_i))$$

$$= \frac{\phi}{w_i} V(\mu_i) \rightarrow V = b''(b'^{-1}) \rightarrow \text{Variance Function}$$

B) Variance function in specific cases:

B.1) $y \sim$ Poisson

we had: $p(y) = \exp\{y_i \cdot \log(\lambda) - \lambda + \log(y_i)\}$

with:
- $\theta = \log(\lambda) \rightarrow e^{\theta} = \lambda \rightarrow -\lambda = -e^{\theta} \Rightarrow b(\theta) = e^{\theta}$
- $b'(\theta) = e^{\theta} \rightarrow (b'(\theta))^{-1} = \log(\theta)$
- $b''(\theta) = e^{\theta}$

$\Rightarrow V(\mu_i) = \exp(\log(\mu_i)) = \mu_i$

B.2) $z \sim$ Binomial

$$p(z) = \exp\left(y \cdot n \cdot \log(p) + (n - yn) \cdot \log(1-p) + \log\left(\frac{n!}{(n-yn)!\,(yn)!}\right)\right)$$

$$= \exp\Big(\underbrace{n}_{1/a(\theta)} \cdot \Big(y \cdot \underbrace{\frac{\log(p)}{\log(1-p)}}_{\theta} + \underbrace{\log(1-p)}_{b(\theta)}\Big)\Big)$$

- $\theta = \dfrac{\log(p)}{\log(1-p)} = \log\left(\dfrac{p}{1-p}\right)$

$$e^{\theta} = \frac{p}{1-p} \rightarrow p = \frac{e^{\theta}}{1+e^{\theta}}$$

$$\Rightarrow b(\theta) = -\log\left(1 - \frac{e^{\theta}}{1+e^{\theta}}\right) = -\log\left(\frac{1}{1+e^{\theta}}\right)$$

$$= \left(-\log\left(1+e^{\theta}\right)\right) = \log\left(1+e^{\theta}\right)$$

- $b'(\theta) = \dfrac{e^{\theta}}{1+e^{\theta}} \rightarrow (b'(\theta))^{-1} = \log\left(\dfrac{\theta}{1-\theta}\right)$

- $b''(\theta) = \dfrac{e^{\theta}}{1+e^{\theta}} - \left(\dfrac{e^{\theta}}{1+e^{\theta}}\right)^2$

$$\Rightarrow V = b''(b'^{-1}(\theta))$$

$$= \frac{e^{\log\left(\frac{\mu}{1-\mu}\right)}}{1+e^{\log\left(\frac{\mu}{1-\mu}\right)}} - \left(\frac{e^{\log\left(\frac{\mu}{1-\mu}\right)}}{1+e^{\log\left(\frac{\mu}{1-\mu}\right)}}\right)^2$$

$$= \frac{\frac{\mu}{1-\mu}}{1+\frac{\mu}{1-\mu}} - \left(\frac{\frac{\mu}{1-\mu}}{1+\frac{\mu}{1+\mu}}\right)^2$$

$$= \mu - \mu^2 = \mu(1-\mu)$$

c) About the LINK FUNCTION (g)

common choice: CANONICAL LINK $\rightarrow g(\mu) = b'^{-1}(\mu)$

Indeed we saw that

$\theta_i = (b')^{-1}(g^{-1}(x_i'\beta))$

hence the canonical link allows for nice simplification

C1 we can use b'^{-1} that we derived in B.1):

$g(\mu) = \log(\mu)$

C2 now using B.2): $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

FITTING GLMS :

A) we can start with different piece:

- $$\frac{\partial \log(L)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left(\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c\left(y_i \mid \frac{\phi}{w_i}\right)\right)$$

$$= \frac{y_i - b'(\theta_i)}{\phi/w_i}$$

- $$\frac{\partial \theta_i}{\partial \mu_i} = \frac{\partial}{\partial \mu_i}\left[(b')^{-1}(\mu_i)\right] = \left((b')^{-1}\right)'(\mu_i)$$

$$= \frac{1}{b''\left((b')^{-1}(\mu_i)\right)}$$

- $$\frac{\partial \mu_i}{\partial \beta} = \frac{\partial}{\partial \beta}\left[g^{-1}(x_i'\beta)\right] = (g^{-1})'(x_i'\beta)\, x_i^T$$

$$= \frac{x_i^T}{g'\left(g^{-1}(x_i'\beta)\right)}$$

So that using the chain rule we have:

$$\frac{\partial \log_i(L)}{\partial \beta} = \frac{\partial \log(L)}{\partial \theta} \cdot \frac{\partial \theta}{\partial \mu} \cdot \frac{\partial \mu}{\partial \beta}$$

$$= \left(\frac{y_i - b'(\theta_i)}{\phi/w_i}\right) \cdot \frac{1}{b''\left((b')^{-1}(\mu_i)\right)} \cdot \frac{x_i'}{g'\left(g^{-1}(x_i'\beta)\right)}$$

$$= \left(\frac{y_i - \mu_i}{\phi / w_i} \right) \cdot \frac{1}{b''(\theta_i)} \cdot \frac{x_i'}{g'(\mu_i)}$$

$$= \frac{w_i (y_i - \mu_i) x_i}{\phi V(\mu_i) g'(\mu_i)}$$

$$\Rightarrow \quad S(\beta, \phi) = \sum^{n} \frac{w_i (y_i - \mu_i) x_i}{\phi V(\mu_i) \cdot g'(\mu_i)}$$

B) we just showed that:

$$S(\beta, \phi) = \sum^{n} \frac{w_i (y_i - \mu_i) x_i}{\phi V(\mu_i) g'(\mu_i)}$$

now recall that under the canonical link we have:

$$g'(\mu_i) = \frac{1}{V(\mu_i)}$$

$$\Rightarrow S(\beta, \phi) = \sum^{n} \frac{w_i (y_i - \mu_i) x_i}{\phi V(\mu_i) / V(\mu_i)} = \sum^{n} \frac{w_i (y_i - \mu_i) x_i}{\phi}$$

c) We now want to fit a GLM

According to the selected model, we have:

$y_i \sim \text{Binom}(N_i, \mu_i)$

so that using what we previously derived, we have:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \Rightarrow g^{-1}(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i\beta}}$$

so that $$s(\beta, \phi) = \sum^{4}\left(y_i - \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)x_i$$

Moreover, recall that a step-size gradient descent that updates β until convergence will work as follows:

$$\beta^{(m+1)} = \beta^{(m)} + \gamma \nabla_\beta \log \mathcal{L}(\beta)$$

For implementation with the whole dataset, see Jupyter notebooks and python code.

In particular, we will use the following information:

- $y_i \sim$ Binomial
- $E(y) = \dfrac{1}{1+e^{-x_i\beta}}$
- $\mathcal{L}() = \prod^{n} \mu_i^{y_i} \cdot (1-\mu_i)^{1-y_i}$

$$\Rightarrow \log \mathcal{L}() = \sum^{n} y_i \cdot \log(\mu_i) + (1-y_i)\cdot \log(1-\mu_i) \log(\mu_i)$$

$$= \sum^{n} \left[y_i \cdot \log \underbrace{\left(\frac{1}{1+e^{-x_i'\beta}} \right)}_{\text{sigmoid function}} + (1-y_i)\cdot \log \left(\frac{e^{-x_i'\beta}}{1+e^{-x_i\beta}} \right)\right.$$

b) We want to get an expression for $\longrightarrow$ $H(\beta,\phi) = \frac{\partial^2}{\partial\beta\partial\beta'} \log \mathcal{L}(\beta,\theta)$

Recall we had: $\frac{\partial}{\partial\beta} \log \mathcal{L}(\beta,\phi) = \sum^n \frac{w_i(y_i-\mu_i)x_i}{\phi}$

So that we can start from that expression and take the partial derivative wrt β':

$$\frac{d}{d\beta^T} \sum^n \frac{w_i(y_i-\mu_i)x_i}{\phi} = \frac{\partial}{\partial\beta'} \sum^n \frac{w_i(y_i - g^{-1}(x_i'\beta))x_i}{\phi}$$

$$= -\frac{1}{\phi} \sum^n w_i \frac{\partial}{\partial\beta'} (g^{-1}(x_i'\beta)) x_i$$

$$= -\frac{1}{\phi} \sum^n w_i (g^{-1})'(x_i'\beta) x_i' x_i$$

$$= -\frac{1}{\phi} \sum^n w_i b''(x_i'\beta) x_i' x_i$$

which can be written in matrix notation for

$$W = \mathrm{diag}\left(\frac{w_i}{\phi} b''(x_i'\beta)\right)$$

$$\Rightarrow H(\beta,\phi) = -X^T W X$$

E) Consider a point $\beta_0 \in \mathbb{R}^p$

we want to approximate $\mathcal{L}(\beta) = \log \mathcal{L}(\beta, \phi)$
with a second order Taylor approximation.
we have:

$$q(\beta, \beta_0) = \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T \Big|_{\beta=\beta_0} (\beta-\beta_0) + \frac{1}{2}(\beta-\beta_0)^T H(\beta_0)(\beta-\beta_0)$$

$$= \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T \Big|_{\beta=\beta_0} (\beta-\beta_0) - \frac{1}{2}(\beta-\beta_0)^T X^T W X(\beta-\beta_0)$$

$$\nabla_\beta \mathcal{L}(\beta)^T = \sum^n \frac{w_i(y_i-\mu_i)x_i'}{\phi}$$

$$= \sum^n \frac{w_i b''(x_i^T\beta)}{\phi} \cdot \frac{(y_i-\mu_i)x_i^T}{b''(x_i^T\beta)}$$

$$= \tilde{z}^T W X$$

$$\left[\dots \frac{y_i-\mu_i}{b''(x_i'\beta)} \dots \right]$$

so that we can also write:

$$q(\beta, \beta_0) = \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T \Big|_{\beta=\beta_0} (\beta-\beta_0) - \frac{1}{2}(\beta-\beta_0)^T X' W X(\beta-\beta_0)$$

$$= \mathcal{L}(\beta_0) + \tilde{z}_{\beta_0} W X(\beta-\beta_0) - \frac{1}{2}(\beta-\beta_0)^T X' W X(\beta-\beta_0)$$

$$= \tilde{z}_{\beta_0}^T W X\beta - \frac{1}{2}\beta^T X^T W X\beta + \beta_0^T X' W X\beta + c^*$$

$$= -\frac{1}{2}\beta^T X' W X\beta + \tilde{y}' W X\beta + c^*$$

$$= -\frac{1}{2}(\tilde{y}-X\beta)^T W(\tilde{y}-X\beta) + c^*$$

with $\tilde{y}^T = \beta_0^T X^T + \tilde{z}_{\beta_0}^T = \beta_0^T X^T + \left[\dots \frac{y_i-\mu_i|_{\beta=\beta_0}}{b''(x_i'\beta_0)} \dots \right]$

7) NEWTON-RAPSON METHOD

The method updates iteratively β^* using the following rule:

$$\beta^* = \beta + H^{-1}_{\mathcal{L}(\hat{\beta})} \nabla \mathcal{L}(\beta)$$

For implementation and results, see notebook & python code.

6) see notebook.