GLM Definition

- 3 components of a GLM
  - A *random* component for the response: $Y_{\mathbf{x}} \sim f(y|\theta_{\mathbf{x}}, \phi)$ where
    - Canonical parameter $\theta_{\mathbf{x}}$: represent the *location*
    - Dispersion parameter ϕ: represent the *scale*
  - A *systematic* (linear) component for the predictors:
    $$\eta_{\mathbf{x}} = \beta_0 + h_1(\mathbf{x})\beta_1 + \cdots + h_{p-1}(\mathbf{x})\beta_{p-1} \equiv X\boldsymbol{\beta}$$
  - A link function g^* such that $\eta_{\mathbf{x}} = g^*(\theta_{\mathbf{x}})$
- Random component
  - Exponential family $f(y|\theta_{\mathbf{x}}, \phi) = \exp\left[\dfrac{y\theta_{\mathbf{x}} - b(\theta_{\mathbf{x}})}{a(\phi)} + c(y, \phi)\right]$
  - Some distributions in exponential family
    - Normal density $N(\mu_{\mathbf{x}}, \sigma^2)$

$$f(y|\theta_{\mathbf{x}}, \phi) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-\mu_{\mathbf{x}})^2}{2\sigma^2}\right] = \exp\left[\frac{y\mu_{\mathbf{x}} - \mu_{\mathbf{x}}^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]$$

$$\Rightarrow \theta_{\mathbf{x}} = \mu_{\mathbf{x}}, \phi = \sigma^2, a(\phi) = \phi, b(\theta_{\mathbf{x}}) = \theta_{\mathbf{x}}^2/2,$$
$$c(y, \phi) = -(y^2/\phi + \log(2\pi\phi))/2$$

- Poisson density $P(\mu_{\mathbf{x}})$

$\theta_x \to E(Y_x)$, $\phi \to Var(Y_x)$

$$f(y|\theta_{\mathbf{x}}, \phi) = \frac{e^{-\mu_{\mathbf{x}}}\mu_{\mathbf{x}}^y}{y!} = \exp\left[y\log\mu_{\mathbf{x}} - \mu_{\mathbf{x}} - \log y!\right]$$

θ_x

$$\Rightarrow \theta_{\mathbf{x}} = \log\mu_{\mathbf{x}}, \phi = 1, a(\phi) = 1, b(\theta_{\mathbf{x}}) = \exp(\theta_{\mathbf{x}}),$$
$$c(y, \phi) = -\log y!$$

If allow ϕ to vary, it becomes dispersion parameter. (Recall overdispersion)

- Binomial density $B(n_{\mathbf{x}}, \mu_{\mathbf{x}})$

related to both $E(Y_x)$ & $Var(Y_x)$ ⇒ so that we can do χ^2 goodness-of-fit test

$$f(y|\theta_{\mathbf{x}}, \phi) = \binom{n_{\mathbf{x}}}{y}\mu_{\mathbf{x}}^y(1-\mu_{\mathbf{x}})^{n_{\mathbf{x}}-y}$$

θ_x

$$= \exp\left[y\log\left(\frac{\mu_{\mathbf{x}}}{1-\mu_{\mathbf{x}}}\right) + n_{\mathbf{x}}\log(1-\mu_{\mathbf{x}}) + \log\binom{n_{\mathbf{x}}}{y}\right]$$

$\mu_x = \frac{e^{\theta_x}}{1+e^{\theta_x}}$

$$\Rightarrow \theta_{\mathbf{x}} = \log\left(\frac{\mu_{\mathbf{x}}}{1-\mu_{\mathbf{x}}}\right), \phi = 1, a(\phi) = 1, c(y, \phi) = \log\binom{n_{\mathbf{x}}}{y}$$
$$b(\theta_{\mathbf{x}}) = -n_{\mathbf{x}}\log(1-\mu_{\mathbf{x}}) = n_{\mathbf{x}}\log(1+\exp(\theta_{\mathbf{x}}))$$

- The Gamma and inverse Gaussian are other (lesser-used) members of the exponential family
- Some other densities such as the negative binomial and Weibull are not members of exponential family, but are sufficiently close so that GLM can be fit with some simple modifications

- ϕ is free in normal while fixed at 1 in Poisson and binomial
- Some authors reserve the term *exponential family* distribution for cases where ϕ is not used (such as Poisson and binomial) while using the term *exponential dispersion family* for cases where it is (such as normal)

➢ Some properties of exponential family

Use $E(\partial\ell/\partial\theta)=0$ & $E(\partial^2\ell/\partial\theta^2)+[E(\partial\ell/\partial\theta)]^2=0$ to find $E(Y)$ & $Var(Y)$. ℓ: log-likelihood function

1. Mean: $E(Y_x) = \mu_x = b'(\theta_x)$
2. Variance: $Var(Y_x) = b''(\theta_x)\, a(\phi)$ (independent of X; depend on X.)

- $E(Y_x)$ does not depend on ϕ
- $Var(Y_x)$ is a product of functions of θ_x and ϕ
- $b''(\theta_x)$ is called the *variance function* and describes how the variance related to the mean
  - In the normal density case, $b''(\theta_x) = 1 \Rightarrow$ variance independent of the mean
  - For other distributions such as Poisson and binomial, $b''(\theta_x)$ is not a constant function $\Rightarrow$ variance depends on the mean
- We can introduce weights by setting $a(\phi) = \phi/w_x$ (scale; no parameter.), where w_x is a *known* weight that varies between observations

- Link function $\eta_x = g^*(\theta_x)$

➢ We now re-write the link function to describe how the mean response, $E(Y_x) = \mu_x$, is linked to η_x, i.e., $\eta_x = g(\mu_x)$ ($\mu_x = b'(\theta_x)$: a function of θ_x.)

$\eta_x \xrightarrow{g} \mu_x \xrightarrow{b'} \theta_x$, $\eta_x \xrightarrow{g^*} \theta_x$, $g^* = g \circ b'$, $g^{*-1} = b'^{-1} \circ g^{-1}$

➢ In principle, g can be any monotone, continuous, and differentiable function

➢ There are some convenient and common choice of g

- Normal (Gaussian) density:
  - Standard choice: $\eta_x = \mu_x$ (identity link) ($Y = X\beta + \varepsilon$; η_x = $E(Y)$)
  - Other choice of g would give (generalization of linear model with non-linear mean) $Y = g^{-1}(X\beta) + \varepsilon$ (Known; $\mu_x = g^{-1}(\eta_x)$)
    Notice that this does not corespond diretly to a tranfrmation on the response as, for example, in Box-Cox type transformation, i.e.
    $Y' = g(Y) = X\beta + \varepsilon$ (like normal after transformation) $\neq$? $Y = g^{-1}(X\beta + \varepsilon) = g^{-1}(X\beta) + \varepsilon^*$
- Poisson density:
  - Identity link $\eta_x = \mu_x$ would possibly cause negative μ_x ($-\infty < \eta_x < \infty$)
  - Standard choice: $\eta_x = \log(\mu_x) \Leftrightarrow \mu_x = \exp(\eta_x) > 0$ (log-linear model)
  - Log link means that additive effect of $\mathbf{x}$ lead to multiplicative effect on μ_x ($\mu_x = e^{\beta_0} \cdot e^{\beta_1 h_1(x)} \cdots$; $\eta_x = \beta_0 + \beta_1 h_1(x) + \cdots + \beta_{p-1} h_{p-1}(x)$)

- Binomial density (treat Y_x/n_x as the response):
  - Standard choice: logit, probit, complementary log-log

➢ Canonical link: choose g such that

$$\eta_x = g(\mu_x) = \theta_x \Rightarrow g^* \text{ is identity function}$$

which means g must satisfy $g(b'(\theta_x)) = \theta_x \Rightarrow g = (b')^{-1}$

- Examples:

Family	Link	Variance function
Gaussian	$\eta_x = \mu_x$	1
Poisson	$\eta_x = \log(\mu_x)$	μ_x
Binomial (proportion)	$\eta_x = \log(\mu_x/(1-\mu_x))$	$\mu_x(1-\mu_x)$
Gamma	$\eta_x = \mu_x^{-1}$	μ_x^2
Inverse Gaussian	$\eta_x = \mu_x^{-2}$	μ_x^3

canonical

$\text{Var}(y) = b''(\theta_x)a(\phi)$

→ $\text{Var}(y)$ is irrelevant to $E(y)$

totally control var(y) ⇒ can do goodness-of-fit test

$b(\theta) = \log(1+e^\theta)$

$\mu = b'(\theta) = \frac{e^\theta}{1+e^\theta}$

$\theta = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

$\text{logit}\left(\frac{e^\theta}{1+e^\theta}\right) = \log\left(\frac{e^\theta/(1+e^\theta)}{1/(1+e^\theta)}\right) = \log(e^\theta) = \theta$

$b'(\theta) = e^\theta/(1+e^\theta)$

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

- If a canonical link is used
  - X^TY is sufficient for β
  - Mathematically and computationally convenient
  - Often physically justified
  - However, it is not required to always use the canonical link and sometimes context may compel another choice

Fitting a GLM ← estimating β

- The log-likelihood for a single observation $(\mathbf{x}_i, y_i)$ is:

check LNp.7-1

$$\log L(\theta_{\mathbf{x}_i}, \phi; y_i) \overset{=}{} w_i \left[\frac{y_i\theta_{\mathbf{x}_i} - b(\theta_{\mathbf{x}_i})}{\phi}\right] + c(y_i, \phi)$$

where $a_i(\phi) = \phi/w_i$ and w_i is a known weight that varies between observations

(w_i not contain any parameters.; cf. $a(\phi)$)

➢ For independent observations $(\mathbf{x}_i, y_i)$, i=1, 2, …, k, the joint log-likelihood is:

$g'(\eta_x) = g'(x\beta)$

$$\sum_{i=1}^{k} \log L(\theta_{\mathbf{x}_i}, \phi; y_i) = \sum_{i=1}^{k} w_i \left[\frac{y_i\theta_{\mathbf{x}_i} - b(\theta_{\mathbf{x}_i})}{\phi}\right] + \sum_{i=1}^{k} c(y_i, \phi) \equiv l(\boldsymbol{\mu}_x, \phi; Y)$$

(with $\boldsymbol{\mu}_x = \mu_x$)

➢ Likelihood approach: the parameters $\boldsymbol{\beta}$, which appear in the link function $X\boldsymbol{\beta} = \eta_x = g^*(\theta_x)$ of a GLM, can be estimated using maximum likelihood --- maximize the joint log-likelihood as a function of $\boldsymbol{\beta}$

- We can maximize the joint log-likelihood *analytically* and find an exact solution for the MLE of $\boldsymbol{\beta}$ --- Gaussian GLM is the only common case where this is possible ⇒ β_{MLE} = least square estor
- Typically, we must use numerical optimization - IRWLS

- Iteratively Re-Weighted Least Squares (IRWLS)
  - Consider the Gaussian linear model $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$
    - Suppose $\text{Var}(y_{\mathbf{x}})=\text{Var}(\varepsilon_{\mathbf{x}}) \propto f(\eta_{\mathbf{x}})$ where $\hat{\eta}_{\mathbf{x}_i}=\mathbf{x}_i^{*T}\hat{\boldsymbol{\beta}}$ ($\mathbf{x}_i^{*}=(1, R_1(x_i), \cdots, R_{p-1}(x_i))$)

assume known, if unknown, use residuals to estimate

residuals: use the WLS to estimate $\boldsymbol{\beta}$ → use the variance structure to estimate $f(\eta_{\mathbf{x}})$ → weights

    - Can use weights w_i where $w_i^{-1}=f(\hat{\eta}_{\mathbf{x}_i})$
    - Since the weights are a function of $\hat{\boldsymbol{\beta}}$, an iterative fitting procedure would be needed
      - set the weights all equal to one $\Rightarrow$ estimate $\hat{\boldsymbol{\beta}} \Rightarrow$ use $\hat{\boldsymbol{\beta}}$ to re-compute the weights $\Rightarrow$ re-estimate $\hat{\boldsymbol{\beta}} \Rightarrow$ repeat until converge
  - similar idea can be applied to fit a GLM

$g(\mu_x)=\eta_x=X\beta$, $\mu_x \leftrightarrow y_x$

    - Roughly speaking, want to regress $g(y_{\mathbf{x}})$ on $\boldsymbol{X}$ with weights inversely proportional to $\text{Var}(g(y_{\mathbf{x}}))$
    - However, $g(y_{\mathbf{x}})$ might not make sense in some cases, e.g., binomial GLM $\Rightarrow$ linearize $g(y_{\mathbf{x}})$ as follows: regress z_x on x

$$g(y_{\mathbf{x}}) \approx g(\mu_{\mathbf{x}})+(y_{\mathbf{x}}-\mu_{\mathbf{x}})g'(\mu_{\mathbf{x}})=\eta_{\mathbf{x}}+(y_{\mathbf{x}}-\mu_{\mathbf{x}})\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}} \equiv z_{\mathbf{x}}$$

Gaussian Linear Model in LNp.7-7 ($\Leftrightarrow$)

- $E(z_{\mathbf{x}})=\eta_{\mathbf{x}} \equiv X\beta$
- $Var(z_{\mathbf{x}})=\left(\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right)^2 Var(y_{\mathbf{x}}) \propto \left(\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right)^2 V(\mu_{\mathbf{x}})$ ($\eta_x=g(\mu_x)$, link; variance function (LNp.7-3); depends on μ_x (mean of y_x), i.e., depends on η_x)
- $\widehat{Var}(z_{\mathbf{x}})=\left(\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right)^2 V(\mu_{\mathbf{x}})\Big|_{\mu_{\mathbf{x}}=\hat{\mu}_{\mathbf{x}}} \equiv \frac{1}{w_{\mathbf{x}}}$

- IRWLS Procedure (estimation)

1. Set initial estimates $\hat{\eta}_{\mathbf{x},0}=g(\hat{\mu}_{\mathbf{x},0})$ and $\hat{\mu}_{\mathbf{x},0}=g^{-1}(\hat{\eta}_{\mathbf{x},0})$ (e.g. $\hat{\mu}_{x,0}=y_x$)
2. Form the "adjusted dependent variable" (x.)

$$z_0=\hat{\eta}_{\mathbf{x},0}+(y_{\mathbf{x}}-\hat{\mu}_{\mathbf{x},0})\left.\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right|_{\hat{\eta}_{\mathbf{x},0}}$$

3. Form the weights (x.)

$$w_0^{-1}=\left.\left(\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right)^2\right|_{\hat{\eta}_{\mathbf{x},0}} V(\hat{\mu}_{\mathbf{x},0})$$

4. Use WLS to re-estimate $\boldsymbol{\beta}$, then get $\hat{\eta}_{\mathbf{x},1}$ and $\hat{\mu}_{\mathbf{x},1}$
5. Iterate steps 2-3-4 until convergence

Q: how much information about the distribution of y been used in the IRWLS?

- Note: the fitting procedure use only $\eta_{\mathbf{x}}=g(\mu_{\mathbf{x}})$ and $V(\mu_{\mathbf{x}})$

$\Rightarrow$ It requires no further knowledge of the distribution of $y_{\mathbf{x}}$

$\Rightarrow$ Quasi-likelihood approach

- Estimates of variance:
  - Comparable to the form used in weighted least squares
  - The weights are now a function of the response for a GLM

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = (X^T W X)^{-1} \hat{\phi}$$

where $\hat{\phi} = X^2/(k-p)$ and X^2 is the Pearson's X^2 statistic

- Example: binomial response (treat the proportion (=count/$n_{\mathbf{x}}$) as the response, not the count)
  - $\eta_{\mathbf{x}} = \log\left(\frac{\mu_{\mathbf{x}}}{1-\mu_{\mathbf{x}}}\right)$
  - $\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}} = \frac{1}{\mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})}$
  - $V(\mu_{\mathbf{x}}) = \frac{\mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})}{n_{\mathbf{x}}}$
  - $w_{\mathbf{x}} = n_{\mathbf{x}}\mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})$
  - $\widehat{Var}(\hat{\boldsymbol{\beta}}) = (X^T W X)^{-1}$
- Some notes about IRWLS:
  - In most cases, the convergence is usually fast
  - If there is a failure to converge, this is often a sign of some problem with the model specification or unusual feature of the data (e.g., data is linearly separable)