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 (Gaussian) 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 $P(\mu_{\mathbf{x}})$

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

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

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

$$\begin{aligned} f(y|\theta_{\mathbf{x}}, \phi) &= \binom{n_{\mathbf{x}}}{y}\mu_{\mathbf{x}}^y(1-\mu_{\mathbf{x}})^{n_{\mathbf{x}}-y} \\ &= \exp\left[y\log\left(\frac{\mu_{\mathbf{x}}}{1-\mu_{\mathbf{x}}}\right) + n_{\mathbf{x}}\log(1-\mu_{\mathbf{x}}) + \log\binom{n_{\mathbf{x}}}{y}\right] \end{aligned}$$

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

- 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

1. Mean: $E(Y_x) = \mu_x = b'(\theta_x)$
2. Variance: $Var(Y_x) = b''(\theta_x)\, a(\phi)$

- $E(Y_x)$ 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 dependents on the mean
- We can introduce weights by setting $a(\phi)=\phi/w_x$, 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., $\eta_x = g(\mu_x)$

➢ 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: $\eta_x = \mu_x$
  - Other choice of g would give

$$\boldsymbol{Y} = g^{-1}(\boldsymbol{X\beta}) + \boldsymbol{\varepsilon}$$

Notice that this does not corespond diretly to a tranfrmation on the response as, for example, in Box-Cox type transformation, i.e.

$$g^{-1}(\boldsymbol{Y}) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

- Poisson density:
  - Identity link $\eta_x = \mu_x$ would possibly cause negative μ_x
  - Standard choice: $\eta_x = \log(\mu_x) \Leftrightarrow \mu_x = \exp(\eta_x) > 0$
  - Log link means that additive effect of $\mathbf{x}$ lead to multiplicative effect on μ_x

- Binomial density (treat $Y_{\mathbf{x}}/n_{\mathbf{x}}$ as the response):
  - Standard choice: logit, probit, complementary log-log

➢ Canonical link: choose g such that

$$\eta_{\mathbf{x}} = g(\mu_{\mathbf{x}}) = \theta_{\mathbf{x}}$$

which means g must satisfy $g(b'(\theta_{\mathbf{x}})) = \theta_{\mathbf{x}}$

- Examples:

Family	Link	Variance function
Gaussian	$\eta_{\mathbf{x}} = \mu_{\mathbf{x}}$	1
Poisson	$\eta_{\mathbf{x}} = \log(\mu_{\mathbf{x}})$	$\mu_{\mathbf{x}}$
Binomial	$\eta_{\mathbf{x}} = \log(\mu_{\mathbf{x}}/(1-\mu_{\mathbf{x}}))$	$\mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})$
Gamma	$\eta_{\mathbf{x}} = \mu_{\mathbf{x}}^{-1}$	$\mu_{\mathbf{x}}^{2}$
Inverse Gaussian	$\eta_{\mathbf{x}} = \mu_{\mathbf{x}}^{-2}$	$\mu_{\mathbf{x}}^{3}$

- If a canonical link is used
  - $X^T Y$ is sufficient for $\boldsymbol{\beta}$
  - 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

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

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

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

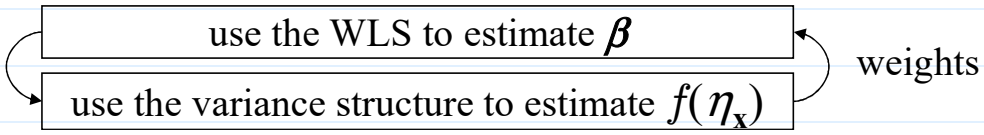
➢ For independent observations $(\mathbf{x}_i, y_i)$, i=1, 2, …, k, the joint log-likelihood is:

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

➢ Likelihood approach: the parameters $\boldsymbol{\beta}$, which appear in the link function $X\boldsymbol{\beta} = \eta_{\mathbf{x}} = g^*(\theta_{\mathbf{x}})$ 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
- Typically, we must use numerical optimization - IRWLS

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



    - 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
    - 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:

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

      - $E(z_{\mathbf{x}}) = \eta_{\mathbf{x}} \equiv X\boldsymbol{\beta}$
      - $Var(z_{\mathbf{x}}) = \left(\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right)^2 Var(y_{\mathbf{x}}) \propto \left(\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right)^2 V(\mu_{\mathbf{x}})$
      - $\widehat{Var}(z_{\mathbf{x}}) = \left(\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right)^2 V(\mu_{\mathbf{x}})\Big|_{\mu_{\mathbf{x}}=\hat{\mu}_{\mathbf{x}}} \equiv \frac{1}{w_{\mathbf{x}}}$
  - IRWLS Procedure
    1. Set initial estimates $\hat{\eta}_{\mathbf{x},0} = g(\hat{\mu}_{\mathbf{x},0})$ and $\hat{\mu}_{\mathbf{x},0} = g^{-1}(\hat{\eta}_{\mathbf{x},0})$
    2. Form the "adjusted dependent variable"
$$z_0 = \hat{\eta}_{\mathbf{x},0} + (y_{\mathbf{x}} - \hat{\mu}_{\mathbf{x},0})\left.\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right|_{\hat{\eta}_{\mathbf{x},0}}$$
    3. Form the weights
$$w_0^{-1} = \left.\left(\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}}\right)^2\right|_{\hat{\eta}_{\mathbf{x},0}} V(\hat{\mu}_{\mathbf{x},0})$$
    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
  - 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

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

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)

- $\eta_{\mathbf{x}} = \log\left(\frac{\mu_{\mathbf{x}}}{1-\mu_{\mathbf{x}}}\right)$
- $\frac{d\eta_{\mathbf{x}}}{d\mu_{\mathbf{x}}} = \frac{1}{\mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})}$
- $V(\mu_{\mathbf{x}}) = \frac{\mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})}{n_{\mathbf{x}}}$
- $w_{\mathbf{x}} = n_{\mathbf{x}}\mu_{\mathbf{x}}(1-\mu_{\mathbf{x}})$
- $\widehat{Var}(\boldsymbol{\beta}) = \left(X^T W X\right)^{-1}$

➢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)

Hypothesis Tests

• null model and saturated model

➢null model: the smallest model we will entertain

- Model for no relation between predictors and response
- Usually, it means we fit a common mean μ for all $y_{\mathbf{x}}$
- For some contingency table models, there will be additional parameters that represent row or column totals or other such constraints $\Rightarrow$ null model has more than one parameter

➢saturated (full) model: the most complex model

- Model in which data is explained exactly
- Typically, k parameters for k data points
- It can be achieved by fitting a sufficiently high-order polynomial or treating quantitative predictors as qualitative predictors or adding enough interactions
- The model tells us no more than the data itself and is usually uninformative

- A statistical model S describes how we partition the data into systematic structure and random variation
  - Null model represents one extreme where the data is represented entirely as random variation
  - Saturated model represents the data as being entirely systematic
  - Model we want usually lie between these two extremes

• Deviance

- **Q**: how to measure discrepancy between observed and fitted y?
- Saturated model gives us a measure of how well *any* model could possibly fit $\Rightarrow$ can consider the difference between the log-likelihood for the saturated and a model S of interest:

$$2(l(Y, \phi; Y) - l(\hat{\boldsymbol{\mu}}, \phi; Y))$$

(which has a rationale from likelihood-ratio test)

- $l(Y, \phi; Y)$: the log-likelihood for the saturated model
- $l(\hat{\boldsymbol{\mu}}, \phi; Y)$: the log-likelihood for the model S

- Provided that the observations are independent and for an exponential family distribution with $a_i(\phi) = \phi/w_i$,

$$2(l(Y, \phi; Y) - l(\hat{\boldsymbol{\mu}}, \phi; Y)) =$$
$$\sum_i 2w_i \left[y_i \left(\tilde{\theta}_{\mathbf{x}_i} - \hat{\theta}_{\mathbf{x}_i} \right) - b(\tilde{\theta}_{\mathbf{x}_i}) + b(\hat{\theta}_{\mathbf{x}_i}) \right] \Big/ \phi \equiv D(Y, \hat{\boldsymbol{\mu}})/\phi$$

where $\tilde{\theta}_{\mathbf{x}}$: the estimates of $\theta_{\mathbf{x}}$ under the saturated model
$\hat{\theta}_{\mathbf{x}}$: the estimates of $\theta_{\mathbf{x}}$ under S

- $D(Y, \hat{\boldsymbol{\mu}})$ is called the *deviance* and $D(Y, \hat{\boldsymbol{\mu}})/\phi$ is called the *scaled deviance*
- Deviance for the common GLM

Family	deviance
Gaussian	$\sum_i (y_i - \hat{\mu}_{\mathbf{x}_i})^2$
Poisson	$2\sum_i [y_i \log(y_i/\hat{\mu}_{\mathbf{x}_i}) - (y_i - \hat{\mu}_{\mathbf{x}_i})]$
Binomial	$2\sum_i [y_i \log(y_i/\hat{\mu}_{\mathbf{x}_i}) + (n - y_i) \log((n - y_i)/(n - \hat{\mu}_{\mathbf{x}_i}))]$
Gamma	$2\sum_i [-\log(y_i/\hat{\mu}_{\mathbf{x}_i}) + (y_i - \hat{\mu}_{\mathbf{x}_i})/\hat{\mu}_{\mathbf{x}_i}]$
Inverse Gaussian	$\sum_i (y_i - \hat{\mu}_{\mathbf{x}_i})^2/(\mu_{\mathbf{x}_i}^2 y_i)$

➢Pearson's X^2 statistic $$X^2 = \sum_i \frac{(y_i - \hat{\mu}_{\mathbf{x}_i})^2}{V(\hat{\mu}_{\mathbf{x}_i})}$$

it is an alternative measure of discrepancy that is sometimes used in replace of the deviance

- Goodness-of-fit test: whether the current model S fit the data
  - ➢Given the model S is correct,
    - $D(Y, \hat{\boldsymbol{\mu}}_S)/\phi \overset{a}{\sim} \chi^2_{df_S}$
    - $X^2 \overset{a}{\sim} \chi^2_{df_S}$
  - ➢For Gaussian GLM, cannot use the test because do not know the value of the dispersion parameter ϕ

    (**Q**: why not replace ϕ by an estimate $\hat{\phi}$?)
  - ➢For the binomial and the Poisson, ϕ=1, so the test is practical
- Difference-in-deviance test: compare two nested models $S \subset L$
  - ➢Given the model S is correct,

$$(D(Y, \hat{\boldsymbol{\mu}}_S) - D(Y, \hat{\boldsymbol{\mu}}_L))/\phi \overset{a}{\sim} \chi^2_{df_S - df_L}$$

➢For the Gaussian model and other models where the dispersion ϕ is *not known*, this chi-square test cannot be directly used

- We can insert an estimate of ϕ and

$$\frac{D(Y, \hat{\boldsymbol{\mu}}_S) - D(Y, \hat{\boldsymbol{\mu}}_L)}{\hat{\phi}_L} \overset{a}{\sim} F_{df_S - df_L, df_L}$$

where $\hat{\phi}_L = X^2_L/df_L$ and X^2_L is the Pearson's X^2 statistic under the model L

- For the Gaussian model, $\hat{\phi}_L = RSS_L/df_L$ and the resulting F-statistic has an exact F distribution under the model S

➢Goodness-of-fit test: L=saturated model

- Notes:
  - ➢The null distribution in the goodness-of-fit and difference-in-deviance test is only asymptotically correct
  - ➢The approximation is better when comparing models than for the goodness of fit statistic
- Wald test for individual β_j:
  - $\hat{\beta}_j/se(\hat{\beta}_j) \overset{a}{\sim} N(0,1)$
  - $\hat{\beta}_j/(se(\hat{\beta}_j)\hat{\phi}) \overset{a}{\sim} t_{df_L}$

- variable selection
  - ➢ stepwise methods
    - Can sequentially (forward or backward or a mix of both) apply difference-in-deviance test to compare nested models (in much the same manner as in standard regression models)
    - The usual concerns about the validity of multiple testing and missing good model carry over
  - ➢ criterion-based methods

$$\text{AIC}_S = \text{Deviance}_S + 2p$$

$$\text{BIC}_S = \text{Deviance}_S + p \log k$$

where p is the number of parameters in the model S and k is the number of covariate classes

- Choose the model with the smallest AIC or BIC
- AIC will tend to pick a larger model than the BIC
- AIC and BIC can be used to compare non-nested model