Generalization to GLM setting: Recall that we are considering distributions of the form:

$$f(y_i|\theta_i,\phi) = \exp\left(\frac{(y_i\theta_i - b(\theta_i))}{\phi} + C(y_i,\phi)\right)$$

In order to generalize this process to the GLM setting, we need some results regarding the function $b(\theta)$.

Since $f(y;\theta,\phi)$ is a density (or discrete distribution function) we have that $\int_\Omega f(y;\theta,\phi)\,d\mu(y) = 1$ where Ω is the sample space (which may be discrete) and $d\mu(y)$ is a dominating measure. [5]

Assuming that we can interchange the order of integration and differentiation,

$$\begin{aligned}\frac{\partial}{\partial\theta}\int_\Omega f(y;\theta,\phi)\,d\mu(y) &= \int_\Omega \frac{\partial}{\partial\theta}\exp(\frac{y\theta - b(\theta)}{\phi} + C)\,d\mu(y)\\ &= \int_\Omega \frac{y - b'(\theta)}{\phi} f(y;\theta,\phi)\,d\mu(y)\\ &= 0\end{aligned}$$

Hence, $\int_\Omega y f(y;\theta,\phi)\,d\mu(y) = \int_\Omega b'(\theta) f(y;\theta,\phi)\,d\mu(y)$ or, $E[y] = b'(\theta)$

Similarly,

$$\frac{\partial^2}{\partial\theta^2}\int_\Omega f(y;\theta,\phi)\,d\mu(y) = \int_\Omega \frac{-b''(\theta)}{\phi} f(y;\theta,\phi)\,d\mu(y) + \int_\Omega \left(\frac{y - b'(\theta)}{\phi}\right)^2 f(y;\theta,\phi)\,d\mu(y) = 0$$

Therefore, $b''(\theta) = \dfrac{\mathrm{Var}(y)}{\phi}$.

Hence, derivatives of l in the GLM setting can be expressed as functions of the mean and variance of y.

Examples:

- Gaussian case: $\theta = \mu$, and $b(\theta) = \mu^2/2$, so $b'(\theta) = \mu$, and $b''(\theta) = 1 = \mathrm{Var}(y)/\sigma^2$.
- Poisson case: $\theta = \log\lambda$, $b(\theta) = e^\theta = \lambda$, $b'(\theta) = e^\theta = \lambda$ and $b''(\theta) = e^\theta = \lambda = \mathrm{Var}(y)$
- Binomial case: $\theta = \log\dfrac{\pi}{1-\pi}$, $b(\theta) = n\log(1+e^\theta)$, $b'(\theta) = \dfrac{ne^\theta}{1+e^\theta} = \pi$ and $b''(\theta) = \dfrac{ne^\theta}{1+e^\theta} - \dfrac{ne^\theta e^\theta}{(1+e^\theta)^2} = n\pi(1-\pi)$
- Hypergeometric case: $\theta = \log\psi$, $b(\theta) = \log(\sum_u K(u)e^{u\theta})$, $b'(\theta) = \dfrac{\sum_u K(u)e^{u\theta}u}{\sum_u K(u)e^{u\theta}} = E[y]$ and $b''(\theta) = \dfrac{\sum_u K(u)e^{u\theta}u^2}{\sum_u K(u)e^{u\theta}} - \dfrac{(\sum_u K(u)e^{u\theta}u)^2}{(\sum_u K(u)e^{u\theta})^2} = E[y^2] - E[y]^2 = \mathrm{Var}(y)$

[5] In the gaussian case, Ω is the real line and $d\mu(y) = dy$. In the discrete case (Poisson, Binomial, *etc.*) we may consider $d\mu(y) = 1$ when y is an integer, and 0 otherwise. In this case the integral is simply the sum over the discrete values of y.

Now let x_i be a vector of covariates β be a vector of parameters. Unlike the Gaussian case, we typically don't want to let $E[y] = x^T\beta$. Instead we model a transformed mean. So, we suppose that

$$g(\mu) = g(E[y_i|\beta]) = x_i^T\beta$$

for some function g called the *link* function.

Every family has a *canonical* link function which is constructed so that $\theta = x_i^T\beta$. I.e.,

$$E[y_i|\beta]) = g^{-1}(\theta) = b'(\theta)$$

We have the following table of canonical link functions:

Family	link function	
Gaussian	identity	$g(\mu) = \mu$
Poisson	log	$g(\lambda) = \log(\lambda)$
Binomial	logit	$g(\pi) = \log(\frac{\pi}{1-\pi})$

(Note that in the hypergeometric case, we typically don't think in terms of sample means, plus, in general the canonical link function does not have a closed form, so we don't usually think about link functions in this case.)

The family (Gaussian, Poisson, binomial, *etc.*), the covariate/parameter space and the link function uniquely determines the model.

With canonical link, the log-likelihood becomes

$$l_i = \frac{1}{\phi}(y_i x_i^T\beta - b(x_i^T\beta)) + C_i$$

so the full log-likelihood becomes

$$\begin{aligned} l &= \frac{1}{\phi}(\sum y_i x_i^T\beta - \sum b(x_i^T\beta)) + C \\ &= \frac{1}{\phi}(Y^T X\beta - \sum b(x_i^T\beta)) + C \end{aligned}$$

Derivative with respect to β:

$$\frac{\partial l}{\partial\beta} = \frac{1}{\phi}(Y^T X - \sum b'(x_i^T\beta)x_i^T)$$

In Gaussian case, $b'(\theta) = \theta$, so this is a linear system. Otherwise it is not, and the solution to $\frac{\partial l}{\partial\beta} = 0$ requires iteration.

Let $B(\beta) = \begin{pmatrix} b'(x_1^T\beta) \\ b'(x_2^T\beta) \\ \vdots \\ b'(x_n^T\beta) \end{pmatrix}$ then $\sum b'(x_i^T\beta)x_i^T = B^T X$ so

$$\frac{\partial l}{\partial\beta} = \frac{1}{\phi}(Y^T X - B^T X) = \frac{1}{\phi}(Y^T - B^T)X$$

The *score* vector is

$$\begin{aligned} U &= \left(\frac{\partial l}{\partial \beta}\right)^T \\ &= \frac{1}{\phi}(X^T(Y-B)) \end{aligned}$$

Note that $B = E[Y]$ when model is correct, so B is the vector of expected values. The MLE, $\hat{\beta}$, solves $X^T(Y-B) = 0$ or $X^TY = X^TB$. Since X is (usually) not invertible ($p < n$) this forces linear combinations of observed values to match linear combinations of fitted values. For many models these linear combinations correspond to marginal totals. Hence, a set of fitted values satisfies the likelihood equations provided that they

1. satisfy the model
2. linear combinations of fitted values match linear combinations of observed values

Example: 2×2 table, no association between exposure and disease

	D+	D-
E+	y_{11}	y_{12}
E-	y_{21}	y_{22}

.

y_{ij} = count in i, j cell, assume that y_{ij} is poisson with mean λ_{ij}.

Model: $\log \lambda_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2j}$ (log-linear)

where $x_{1i} = \begin{cases} 1 & \text{exposed (i=1)} \\ 0 & \text{not-exposed (i=2)} \end{cases}$ and $x_{2j} = \begin{cases} 1 & \text{diseased (j=1)} \\ 0 & \text{non-diseased (j=2)} \end{cases}$

(Note that $\log \psi = \log \lambda_{11} - \log \lambda_{12} - \log \lambda_{21} + \log \lambda_{22} = 0$)

$$Y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad X^TY = \begin{bmatrix} y_{11} + y_{12} + y_{21} + y_{22} \\ y_{11} + y_{12} \\ y_{11} + y_{21} \end{bmatrix}$$ So Score equations force the corect marginal totals,

$$\begin{aligned} y_{11} + y_{12} + y_{21} + y_{22} &= E[y_{11}] + E[y_{12}] + E[y_{21}] + E[y_{22}] &&= e^{\beta_0+\beta_1+\beta_2} + e^{\beta_0+\beta_1} + e^{\beta_0+\beta_2} + e^{\beta_0} \\ y_{11} + y_{12} &= E[y_{11}] + E[y_{12}] &&= e^{\beta_0+\beta_1+\beta_2} + e^{\beta_0+\beta_1} \\ y_{11} + y_{21} &= E[y_{11}] + E[y_{21}] &&= e^{\beta_0+\beta_1+\beta_2} + e^{\beta_0+\beta_2} \end{aligned}$$

The Model forces O.R. = 1, so

$$\hat{\lambda}_{ij} = \frac{(y_{1i} + y_{2i})(y_{1j} + y_{2j})}{y_{11} + y_{12} + y_{21} + y_{22}}$$

satisfies both the model and the marginal totals.

In general, $X^T(Y-B) = 0$ is a set of non-linear equations. We may solve them via Newton-Raphson. *I.e.*, compute

$$\frac{\partial U}{\partial \beta} = -\frac{1}{\phi} X^T \frac{\partial B}{\partial \beta}$$

and $\dfrac{\partial B}{\partial \beta} =$ Matrix with ij entry:

$$\frac{\partial}{\partial \beta_j} b'(x_i^T\beta) = b''(x_i^T\beta)x_{ij}$$

So

$$\begin{aligned}\frac{\partial B}{\partial \beta} &= \begin{pmatrix} b''(x_1^T\beta)x_{11} & b''(x_1^T\beta)x_{12} & \cdots & b''(x_1^T\beta)x_{1p} \\ b''(x_2^T\beta)x_{21} & b''(x_2^T\beta)x_{22} & \cdots & b''(x_2^T\beta)x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ b''(x_n^T\beta)x_{n1} & b''(x_n^T\beta)x_{n2} & \cdots & b''(x_n^T\beta)x_{np} \end{pmatrix} \\ &= WX\end{aligned}$$

where

$$\begin{aligned}W &= \begin{pmatrix} b''(x_1^T\beta) & 0 & \cdots & 0 \\ 0 & b''(x_2^T\beta) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & b''(x_n^T\beta) \end{pmatrix} \\ &= \frac{1}{\phi}\mathrm{Cov}(Y)\end{aligned}$$

So

$$\frac{\partial U}{\partial \beta} = -\frac{1}{\phi}X^TWX \quad \text{(In Gaussian Case, } W = I\text{)}$$

$E[\partial U/\partial \beta]$ is the *Fisher Information Matrix*. (When we use the canonical link, it does not depend on Y, so we can drop the expectation.)

Now consider the Taylor series expansion of U, around some initial value $\beta^{(0)}$

$$U(\beta) \approx U(\beta^{(0)}) + \frac{\partial U}{\partial \beta}(\beta - \beta^{(0)})$$

or,

$$\frac{1}{\phi}X^T(Y - B(\beta)) \approx \frac{1}{\phi}X^T(Y - B(\beta^{(0)})) - \frac{1}{\phi}X^TW(\beta^{(0)})X(\beta - \beta^{(0)})$$

Since the MLE solves the LHS = 0, we solve the RHS = 0:

$$X^T(Y - B(\beta^{(0)})) = X^TW(\beta^{(0)})X(\beta - \beta^{(0)})$$

or

$$\beta^{(1)} = \beta^{(0)} + (X^TW(\beta^{(0)})X)^{-1}X^T(Y - B(\beta^{(0)}))$$

(In Gaussian case, $W = I$ and $B = X\beta$, so

$$\begin{aligned}\beta^{(1)} &= \beta^{(0)} + (X^TX)^{-1}X^T(Y - X\beta^{(0)}) \\ &= (X^TX)^{-1}X^TY\end{aligned}$$

and we get the solution in one step.) Otherwise, iterate,

$$\beta^{(i+1)} = \beta^{(i)} + (X^T W(\beta^{(i)})X)^{-1} X^T (Y - B(\beta^{(i)}))$$

until convergence is acheived.

If non-canonical link is used, $\dfrac{\partial U}{\partial \beta}$ depends on Y (via $\sum y_i \dfrac{\partial^2 \theta_i}{\partial \beta^2}$). In this case, we may use $E[\frac{\partial U}{\partial \beta}]$ ($-$Fisher information matrix) in place of $\frac{\partial U}{\partial \beta}$. (This is what the function `glm` in Splus does (at least in version 3), for example). This called *Fisher Scoring*. Note that when the canonical link is used, Fisher Scoring is equivalent to Newton-Raphson. Alternatively, $\hat{\beta}$ can be estimated *via iteratively re-weighted least squares* (This is what SAS `PROC LOGISTIC` and the `glm` function in R do.).