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

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

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 \(\Omega\) is the sample space (which may be discrete) and \(d\mu(y)\) is a dominating measure.  

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

\[
\frac{\partial}{\partial \theta} \int_{\Omega} f(y; \theta, \phi) \, d\mu(y) = \int_{\Omega} \frac{\partial}{\partial \theta} \exp \left( \frac{y \theta - b(\theta)}{\phi} + C \right) \, d\mu(y)
\]

\[
= \int_{\Omega} \frac{y - b'(\theta)}{\phi} f(y; \theta, \phi) \, d\mu(y)
\]

\[
= 0
\]

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, \(b''(\theta) = \frac{\text{Var}(y)}{\phi}\).

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 \(b''(\theta) = 1 = \text{Var}(y)/\sigma^2\).

- Poisson case: \(\theta = \log \lambda\), \(b(\theta) = e^\theta = \lambda\), \(b'(\theta) = e^\theta = \lambda\) and \(b''(\theta) = e^\theta = \lambda = \text{Var}(y)\)

- Binomial case: \(\theta = \log \frac{\pi}{1 - \pi}\), \(b(\theta) = n \log(1 + e^\theta)\), \(b'(\theta) = \frac{ne^\theta}{1 + e^\theta} = \pi\) and \(b''(\theta) = \frac{ne^\theta}{1 + e^\theta} - \frac{ne^\theta e^\theta}{(1 + e^\theta)^2} = n\pi(1 - \pi)\).

- Hypergeometric case: \(\theta = \log \psi\), \(b(\theta) = \log (\sum_u K(u)e^{u\theta})\), \(b'(\theta) = \frac{\sum_u K(u) e^{u\theta} u}{\sum_u K(u) e^{u\theta}} = E[y]\) and \(b''(\theta) = \frac{\sum_u K(u) e^{u\theta} u^2}{\sum_u K(u) e^{u\theta}} - \frac{(\sum_u K(u) e^{u\theta} u)^2}{(\sum_u K(u) e^{u\theta})^2} = E[y^2] - E[y]^2 = \text{Var}(y)\)

\(^5\)In the gaussian case, \(\Omega\) 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 \( \beta \) be a vector of parameters. Unlike the Gaussian case, we typically don’t want to let \( E[y] = x_i^T \beta \). 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:

<table>
<thead>
<tr>
<th>Family</th>
<th>link function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>identity</td>
</tr>
<tr>
<td>Poisson</td>
<td>log</td>
</tr>
<tr>
<td>Binomial</td>
<td>logit</td>
</tr>
</tbody>
</table>

(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
\[
l = \frac{1}{\phi}\left(\sum y_i x_i^T \beta - \sum b(x_i^T \beta)\right) + C
\]
\[
= \frac{1}{\phi}(Y^T X \beta - \sum b(x_i^T \beta)) + C
\]

Derivative with respect to \( \beta \):
\[
\frac{\partial l}{\partial \beta} = \frac{1}{\phi}(Y^T X - \sum b'(x_i^T \beta)x_i^T)
\]
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

\[ \mathbf{U} = \left( \frac{\partial l}{\partial \beta} \right)^T = \frac{1}{\phi} (\mathbf{X}^T (\mathbf{Y} - \mathbf{B})) \]

Note that \( \mathbf{B} = E[\mathbf{Y}] \) when model is correct, so \( \mathbf{B} \) is the vector of expected values. The MLE, \( \hat{\beta} \), solves \( \mathbf{X}^T (\mathbf{Y} - \mathbf{B}) = 0 \) or \( \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{B} \). Since \( \mathbf{X} \) 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 \( \times \) 2 table, no association between exposure and disease

<table>
<thead>
<tr>
<th></th>
<th>E+</th>
<th>D+</th>
<th>E-</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( y_{11} )</td>
<td>( y_{12} )</td>
<td>( y_{21} )</td>
<td>( y_{22} )</td>
</tr>
</tbody>
</table>

\( y_{ij} \) = count in \( i,j \) cell, assume that \( y_{ij} \) is poisson with mean \( \lambda_{ij} \).

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

\[ \mathbf{Y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \]

and \( \mathbf{X}^T \mathbf{Y} = \begin{bmatrix} y_{11} + y_{12} + y_{21} + y_{22} \\ y_{11} + y_{12} \\ y_{11} + y_{21} \end{bmatrix} \)

So Score equations force the correct marginal totals,

\[ \frac{y_{11} + y_{12} + y_{21} + y_{22}}{y_{11} + y_{12}} = E[y_{11}] + E[y_{12}] + E[y_{21}] + E[y_{22}] = e^{\beta_0 + \beta_1 + \beta_2}, \quad \frac{y_{11} + y_{12}}{y_{11} + y_{21}} = e^{\beta_0 + \beta_1 + \beta_2} \]

The Model forces O.R. = 1, so

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

satisfies both the model and the marginal totals.

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

\[ \frac{\partial \mathbf{U}}{\partial \beta} = -\frac{1}{\phi} \mathbf{X}^T \frac{\partial \mathbf{B}}{\partial \beta} \]
and $\frac{\partial B}{\partial \beta} = \text{Matrix with } i j \text{ entry:}$

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

So

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

where

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

So

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

$E[\partial U/\partial \beta]$ 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,

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

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

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

or

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

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

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

$$= (X^T X)^{-1} X^T Y$$
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 achieved.

If non-canonical link is used, \( \frac{\partial U}{\partial \beta} \) depends on \( Y \) (via \( \sum y_i \frac{\partial^2 \theta_i}{\partial \beta^2} \)). 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 \texttt{glm} in Splus does (at least in version 3), for example). This called \textit{Fisher Scoring}. Note that when the canonical link is used, Fisher Scoring is equivalent to Newton-Raphson. Alternatively, \( \hat{\beta} \) can be estimated \textit{via iteratively re-weighted least squares} (This is what SAS PROC \texttt{LOGISTIC} and the \texttt{glm} function in R do.).