

Generalized Linear Models (GLM)

Ludovic Stourm

June 25, 2025

1 Setup

1.1 Generalized linear model

Model:

$$Y_n \sim \mathcal{F}(V_n) \quad \text{where } V_n = X_n \beta \quad (1)$$

Likelihood:

$$\begin{aligned} LL(\beta) &= \sum_n f(V_n) \\ &= \boxed{\mathbb{1}_N \cdot \mathbf{f}} \quad \text{where } f_n = f(V_n) \end{aligned}$$

Gradient:

$$\begin{aligned} \frac{\partial LL}{\partial \beta_l} &= \sum_n X_{nl} \frac{\partial f(V_n)}{\partial V_n} \\ \nabla LL(\beta) &= \boxed{\mathbf{X}' \mathbf{g}} \quad \text{where } g_n = \frac{\partial f(V_n)}{\partial V_n} \end{aligned} \quad (2)$$

Hessian:

$$\begin{aligned} \frac{\partial^2 LL}{\partial \beta_l \partial \beta} &= \sum_n X_{nl} X_{nl'} \frac{\partial^2 f(V_n)}{\partial V_n^2} \\ \mathbf{H}(\beta) &= \boxed{\mathbf{X}' [\mathbf{X} \odot (\mathbf{h} \cdot \mathbb{1}_P)]} \quad \text{where } h_n = \frac{\partial^2 f(V_n)}{\partial V_n^2} \end{aligned}$$

where \mathbf{X} is a matrix of dimensions $[N \times K]$, β is a vector of dimensions $[K \times 1]$, \mathbf{f} , \mathbf{g} and \mathbf{h} are vectors of dimensions $[N \times 1]$, $\mathbb{1}_Q$ is a vector of ones of dimension $[1 \times Q]$, and \odot represents the Hadamard product (term-by-term multiplication).

1.2 Model 1: Linear model

Model:

$$Y_n \sim N(V_n, \sigma^2) \quad (3)$$

Link function and derivatives:

$$\begin{cases} f_n &= -\frac{1}{2\sigma^2}(y_n - V_n)^2 \\ g_n &= \frac{1}{\sigma^2}(y_n - V_n) \\ h_n &= -\frac{1}{\sigma^2} \end{cases} \quad (4)$$

The gradient is equal to zero when:

$$\begin{aligned} \sum_n X_{nl}(y_n - X_n\beta) &= 0 \quad \forall l \\ \Rightarrow \sum_n X_{nl}y_n &= \sum_n X_{nl}X_n\beta \quad \forall l \\ \Rightarrow \mathbf{X}'\mathbf{Y} &= \mathbf{X}'(\mathbf{X}\beta) \\ \Rightarrow \beta &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \end{aligned} \quad (5)$$

1.3 Model 2: Poisson count model / Exponential duration model

Model:

$$\begin{aligned} Y_n &\sim \text{Poisson}(\lambda_n) \\ \log(\lambda_n) &= C_n + V_n \end{aligned} \quad (6)$$

Link function and derivatives:

$$\begin{cases} f_n &= -\lambda_n + y_n V_n + K_n \\ &\text{where } K_n = y_n C_n - \log(y_n!) \text{ does not depend on } \beta \\ g_n &= -\lambda_n + y_n \\ h_n &= -\lambda_n \end{cases} \quad (7)$$

Remarks:

- This subsumes a Poisson count process such that Y_n counts the number of events arising at rate μ_n within a time interval of duration D_n . In that case:

$$\begin{aligned} Y_n &\sim \text{Poisson}(D_n\mu_n) \\ \log(\mu_n) &= A_n + V_n \\ \rightarrow \text{Define: } \lambda_n &= D_n\mu_n \\ \text{and: } C_n &= \log(D_n) + A_n \end{aligned} \quad (8)$$

- By duality of the Poisson process and the Exponential duration model, this subsumes a duration model with constant hazard rate and potential truncation (up to a normalizing

constant). In that case:

$$\begin{aligned}
T_n^* &\sim \text{Exponential}(\mu_n) \\
\log(\mu_n) &= C_n + V_n \\
T_n &= \min\{T_n^*, T_{\max}\} \\
\rightarrow \text{Define: } Y_n &= \mathbb{1}\{T_n = T_n^*\} \text{ (indicates whether the event occurs within } [0, T_{\max}]) \\
\text{and: } \lambda_n &= \mu_n T_n \\
\text{Then: } L_n &= \mu_n^{y_n} e^{-\mu_n T_n} = (1/T_n)^{y_n} \lambda_n^{y_n} e^{-\lambda_n} \\
f_n &= -\lambda_n + y_n V_n + K_n^* \\
&\text{where } K_n^* = y_n C_n \text{ does not depend on } \beta
\end{aligned} \tag{9}$$

Thus, only the constant K_n^* is different.

1.4 Model 3: Binomial logit / Logistic regression

Model:

$$\begin{aligned}
Y_n &\sim \text{Multinomial}(M_n, p_n) \\
p_n &= 1 / (1 + \exp[-V_n])
\end{aligned} \tag{10}$$

Link function and derivatives:

$$\begin{cases}
f_n &= y_n V_n - M_n \log(1 + \exp[V_n]) + K_n \\
&= y_n \log(p_n) + (M_n - y_n) \log(1 - p_n) + K_n \\
&\text{where } K_n = \log(M_n! / [y_n!(M_n - y_n)!]) \text{ does not depend on } \beta \\
g_n &= y_n - M_n p_n \\
h_n &= -M_n p_n (1 - p_n)
\end{cases} \tag{11}$$

2 Estimation by Maximum Likelihood

Newton-Raphson:

$$\beta^{(i+1)} \leftarrow \beta^{(i)} - [H(\beta)]^{-1} \nabla_{LL}(\beta) \tag{12}$$

Here are the things that need to be computed efficiently to estimate the model:

- $\mathbf{V} = \mathbf{X}\beta$
- $\nabla_{LL}(\beta) = \mathbf{X}'\mathbf{g}$
- $\mathbf{X}'[\mathbf{X} \odot (\mathbf{h} \cdot \mathbb{1}_P)]$
- $\mathbf{X}'\mathbf{X}$ (special case of the previous line, where \mathbf{h} is a column of ones.

2.1 Case with two dimensions of variation

The matrix of covariates \mathbf{X} is of dimensions $[N \times P]$. We consider the case when the data varies across two dimensions with subscripts (i, j) , such that the data is balanced in the sense that all combinations of (i, j) appear exactly once (in which case $N = I \times J$). Furthermore, some of the covariates X vary only according to i (and are constant across j), and some covariates vary only according to j (and are constant across i). In that case, we can avoid the full expansion of matrix X (which may require a large chunk of memory), and perform computations more efficiently.

We split the matrix of covariates X_{ij} into the three corresponding groups to obtain matrices $X_{ij}^{(1)}$, $X_i^{(2)}$, $X_j^{(3)}$, and we denote by $\beta^{(1)}, \beta^{(2)}, \beta^{(3)}$ the corresponding subvectors of β .

- To compute $\mathbf{V} = \mathbf{X}\beta$:

$$\begin{aligned} V_{ij} &= X_{ij}\beta \\ &= X_{ij}^{(1)}\beta^{(1)} + X_i^{(2)}\beta^{(2)} + X_j^{(3)}\beta^{(3)} \end{aligned} \quad (13)$$

- To compute $\nabla LL(\beta) = \mathbf{X}'\mathbf{g}$:

$$\begin{aligned} \nabla LL^{(1)} &= \sum_i \sum_j g_{ij} X_{ij}^{(1)} \\ \nabla LL^{(2)} &= \sum_i \left[\sum_j g_{ij} \right] X_i^{(2)} = \sum_i g_i^{(2)} X_i^{(2)} \quad \text{where } g_i^{(2)} = \sum_j g_{ij} \\ \nabla LL^{(3)} &= \sum_j \left[\sum_i g_{ij} \right] X_j^{(3)} = \sum_j g_j^{(3)} X_j^{(3)} \quad \text{where } g_j^{(3)} = \sum_i g_{ij} \end{aligned} \quad (14)$$

- To compute $\mathbf{X}'[\mathbf{X} \odot (\mathbf{h} \cdot \mathbb{1}_P)]$:

$$\begin{aligned} H^{(1,1)} &= \sum_i \sum_j h_{ij} X_{ij}^{(1)} X_{ij}^{(1)} \\ H^{(1,2)} &= \sum_i \left[\sum_j h_{ij} X_{ij}^{(1)} \right] X_i^{(2)} \\ H^{(1,3)} &= \sum_j \left[\sum_i h_{ij} X_{ij}^{(1)} \right] X_j^{(3)} \\ H^{(2,2)} &= \sum_i \left[\sum_j h_{ij} \right] X_i^{(2)} X_i^{(2)} \\ H^{(2,3)} &= \sum_j \left[\sum_i h_{ij} X_i^{(2)} \right] X_j^{(3)} \\ H^{(3,3)} &= \sum_j \left[\sum_i h_{ij} \right] X_j^{(3)} X_j^{(3)} \\ H^{(i,j)} &= H^{(j,i)} \quad \text{if } i > j \end{aligned}$$

2.2 Case with three dimensions of variation

Similarly, let us now consider the case when the data varies across three dimensions with subscripts (i, j, k) , such that the data is balanced in the sense that all combinations of (i, j, k) appear exactly once (in which case $N = I \times J \times K$). We similarly split the matrix of covariates X_{ijk} into six groups, as a function of their dimension(s) of variations, to obtain matrices $X_{ijk}^{(1)}, X_{ij}^{(2)}, X_{ik}^{(3)}, X_{jk}^{(4)}, X_i^{(5)}, X_j^{(6)}, X_k^{(7)}$, and we denote by $\beta^{(1)}, \beta^{(2)}, \beta^{(3)}, \beta^{(4)}, \beta^{(5)}, \beta^{(6)}, \beta^{(7)}$ the corresponding subvectors of β .

- To compute $\mathbf{V} = \mathbf{X}\beta$:

$$\begin{aligned} V_{ijk} &= X_{ijk}\beta \\ &= X_{ijk}^{(1)}\beta^{(1)} + X_{ij}^{(2)}\beta^{(2)} + X_{ik}^{(3)}\beta^{(3)} + X_{jk}^{(4)}\beta^{(4)} + X_i^{(5)}\beta^{(5)} + X_j^{(6)}\beta^{(6)} + X_k^{(7)}\beta^{(7)} \end{aligned} \quad (15)$$

- To compute $\nabla LL(\beta) = \mathbf{X}'\mathbf{g}$:

$$\begin{aligned} \nabla LL^{(1)} &= \sum_i \sum_j \sum_k g_{ijk} X_{ijk}^{(1)} \\ \nabla LL^{(2)} &= \sum_i \sum_j \left[\sum_k g_{ijk} \right] X_{ij}^{(2)} = \sum_i \sum_j g_{ij}^{(2)} X_{ij}^{(2)} \quad \text{where } g_{ij}^{(2)} = \sum_k g_{ijk} \\ \nabla LL^{(3)} &= \sum_k \sum_i \left[\sum_j g_{ijk} \right] X_{ik}^{(3)} = \sum_k \sum_i g_{ik}^{(3)} X_{ik}^{(3)} \quad \text{where } g_{ik}^{(3)} = \sum_j g_{ijk} \\ \nabla LL^{(4)} &= \sum_k \sum_j \left[\sum_i g_{ijk} \right] X_{jk}^{(4)} = \sum_k \sum_j g_{jk}^{(4)} X_{jk}^{(4)} \quad \text{where } g_{jk}^{(4)} = \sum_i g_{ijk} \\ \nabla LL^{(5)} &= \sum_i \left[\sum_k \sum_j g_{ijk} \right] X_i^{(5)} = \sum_i g_i^{(5)} X_i^{(5)} \quad \text{where } g_i^{(5)} = \sum_{j,k} g_{ijk} \\ \nabla LL^{(6)} &= \sum_j \left[\sum_k \sum_i g_{ijk} \right] X_j^{(6)} = \sum_j g_j^{(6)} X_j^{(6)} \quad \text{where } g_j^{(6)} = \sum_{i,k} g_{ijk} \\ \nabla LL^{(7)} &= \sum_k \left[\sum_i \sum_j g_{ijk} \right] X_k^{(7)} = \sum_k g_k^{(7)} X_k^{(7)} \quad \text{where } g_k^{(7)} = \sum_{i,j} g_{ijk} \end{aligned} \quad (16)$$

Thus, we first compute $g_{ij}^{(2)}, g_{ik}^{(3)}, g_{jk}^{(4)}, g_i^{(5)}, g_j^{(6)}, g_k^{(7)}$. Then, we multiply them with the corresponding matrices $X_{ij}^{(2)}, X_{ik}^{(3)}, X_{jk}^{(4)}, X_i^{(5)}, X_j^{(6)}, X_k^{(7)}$. We combine all gradient sub-vectors together to obtain the gradient with respect to the full vector of parameters β . Following this process allows us to avoid the full expansion of X at the (i, j, k) -level.

- To compute $\mathbf{X}'[\mathbf{X} \odot (\mathbf{h} \cdot \mathbb{1}_P)]:$

$$\begin{aligned}
H^{(1,1)} &= \sum_i \sum_j \sum_k h_{ijk} X_{ijk}^{(1)} X_{ijk}^{(1)} \\
H^{(1,2)} &= \sum_i \sum_j \left[\sum_k h_{ijk} X_{ijk}^{(1)} \right] X_{ij}^{(2)} \\
H^{(1,3)} &= \sum_i \sum_k \left[\sum_j h_{ijk} X_{ijk}^{(1)} \right] X_{ik}^{(3)} \\
H^{(1,4)} &= \sum_j \sum_k \left[\sum_i h_{ijk} X_{ijk}^{(1)} \right] X_{jk}^{(4)} \\
H^{(1,5)} &= \sum_i \left[\sum_j \sum_k h_{ijk} X_{ijk}^{(1)} \right] X_i^{(5)} \\
H^{(1,6)} &= \sum_j \left[\sum_i \sum_k h_{ijk} X_{ijk}^{(1)} \right] X_j^{(6)} \\
H^{(1,7)} &= \sum_k \left[\sum_i \sum_j h_{ijk} X_{ijk}^{(1)} \right] X_k^{(7)} \\
H^{(2,2)} &= \sum_i \sum_j \left[\sum_k h_{ijk} \right] X_{ij}^{(2)} X_{ij}^{(2)} \\
H^{(2,3)} &= \sum_k \sum_i \left[\sum_j h_{ijk} X_{ij}^{(2)} \right] X_{ik}^{(3)} \\
H^{(2,4)} &= \sum_k \sum_j \left[\sum_i h_{ijk} X_{ij}^{(2)} \right] X_{jk}^{(4)} \\
H^{(2,5)} &= \sum_i \left[\sum_j \left(\sum_k h_{ijk} \right) X_{ij}^{(2)} \right] X_i^{(5)} \\
H^{(2,6)} &= \sum_j \left[\sum_i \left(\sum_k h_{ijk} \right) X_{ij}^{(2)} \right] X_j^{(6)} \\
H^{(2,7)} &= \sum_k \left[\sum_i \sum_j h_{ijk} X_{ij}^{(2)} \right] X_k^{(7)} \\
H^{(3,3)} &= \sum_k \sum_i \left[\sum_j h_{ijk} \right] X_{ik}^{(3)} X_{ik}^{(3)} \\
H^{(3,4)} &= \sum_k \sum_j \left[\sum_i h_{ijk} X_{ik}^{(3)} \right] X_{jk}^{(4)} \\
H^{(3,5)} &= \sum_i \left[\sum_k \left(\sum_j h_{ijk} \right) X_{ik}^{(3)} \right] X_i^{(5)} \\
H^{(3,6)} &= \sum_j \left[\sum_k \sum_i h_{ijk} X_{ik}^{(3)} \right] X_j^{(6)} \\
H^{(3,7)} &= \sum_k \left[\sum_i \left(\sum_j h_{ijk} \right) X_{ik}^{(3)} \right] X_k^{(7)} \\
H^{(4,4)} &= \sum_k \sum_j \left[\sum_i h_{ijk} \right] X_{jk}^{(4)} X_{jk}^{(4)} \\
H^{(4,5)} &= \sum_i \left[\sum_k \sum_j h_{ijk} X_{jk}^{(4)} \right] X_i^{(5)} \\
H^{(4,6)} &= \sum_j \left[\sum_k \left(\sum_i h_{ijk} \right) X_{jk}^{(4)} \right] X_j^{(6)} \\
H^{(4,7)} &= \sum_k \left[\sum_j \left(\sum_i h_{ijk} \right) X_{jk}^{(4)} \right] X_k^{(7)} \\
H^{(5,5)} &= \sum_i \left[\sum_k \sum_j h_{ijk} \right] X_i^{(5)} X_i^{(5)} \\
H^{(5,6)} &= \sum_j \left[\sum_i \left(\sum_k h_{ijk} \right) X_i^{(5)} \right] X_j^{(6)} \\
H^{(5,7)} &= \sum_i \left[\sum_k \left(\sum_j h_{ijk} \right) X_k^{(7)} \right] X_i^{(5)} \\
H^{(6,6)} &= \sum_j \left[\sum_k \sum_i h_{ijk} \right] X_j^{(6)} X_j^{(6)} \\
H^{(6,7)} &= \sum_j \left[\sum_k \left(\sum_i h_{ijk} \right) X_k^{(7)} \right] X_j^{(6)} \\
H^{(7,7)} &= \sum_k \left[\sum_i \sum_j h_{ijk} \right] X_k^{(7)} X_k^{(7)} \\
H^{(i,j)} &= H^{(j,i)} \quad \text{if } i > j
\end{aligned}$$