Generalized Linear Models (GLM)

Ludovic Stourm

July 1, 2025

1 Univariate generalized linear model

1.1 General setup

Model:

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

where Y_n and V_n are scalars, $\mathcal{F}$ is some probability distribution, X_n is a $[1 \times L]$ vector of observables, and β is a $[L \times 1]$ vector of parameters to estimate.

$$\begin{aligned}
&\text{Likelihood:} && \\
&LL(\beta) && = \sum_n f(V_n) \\
& && = \boxed{\mathbb{1}_N \cdot \mathbf{f}} && \text{where } f_n = f(V_n) \\
&\text{Gradient:} && \\
&\frac{\partial LL}{\partial \beta_l} && = \sum_n X_{nl} \frac{\partial f(V_n)}{\partial V_n} \\
&\nabla LL(\beta) && = \boxed{\mathbf{X}'\mathbf{g}} && \text{where } g_n = \frac{\partial f(V_n)}{\partial V_n} \\
&\text{Hessian:} && \\
&\frac{\partial^2 LL}{\partial \beta_l \partial \beta} && = \sum_n X_{nl} X_{nl'} \frac{\partial^2 f(V_n)}{\partial V_n^2} \\
&\mathbf{H}(\beta) && = \boxed{\mathbf{X}'\left[\mathbf{X} \odot (\mathbf{h} \cdot \mathbb{1}_P)\right]} && \text{where } h_n = \frac{\partial^2 f(V_n)}{\partial V_n^2}
\end{aligned} \tag{2}$$

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 Examples

1.2.1 Model 1: Linear model

Model:

$$Y_n \sim N(V_n, \sigma^2) \tag{3}$$

Link function and derivatives:

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

The gradient is equal to zero when:

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

1.2.2 Model 2: Poisson count model / Exponential duration model

Model:

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

Link function and derivatives:

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

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:

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

- 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:

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

Thus, only the constant K_n^* is different.

1.2.3 Model 3: Binomial logit / Logistic regression

Model:

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

Link function and derivatives:

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

1.3 Estimation by Maximum Likelihood

Newton-Raphson:

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

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

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

1.3.1 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 $\mathbf{V} = \mathbf{X}\beta$:

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

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

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

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

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

1.3.2 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 $\mathbf{V} = \mathbf{X}\beta$:

$$\begin{aligned} V_{ijk} &= X_{ijk}\beta \\ &= X_{ijk}^{(1)}\beta^{(1)} + X_{ij}^{(2)}\beta^{(2)} + X_{ik}^{(3)}\beta^{(3)} + X_{jk}^{(4)}\beta^{(4)} + X_{i}^{(5)}\beta^{(5)} + X_{j}^{(6)}\beta^{(6)} + X_{k}^{(7)}\beta^{(7)} \end{aligned} \tag{15}$$

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

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

  Thus, we first compute $g_{ijk}^{(1)}$, $g_{ij}^{(2)}$, $g_{ik}^{(3)}$, $g_{jk}^{(4)}$, $g_{i}^{(5)}$, $g_{j}^{(6)}$, $g_{k}^{(7)}$. 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 $\mathbf{X}' \left[\mathbf{X} \odot (\mathbf{h} \cdot \mathbb{1}_P) \right]$:

$$H^{(1,1)} = \sum_i \sum_j \sum_k h_{ijk} X^{(1)}_{ijk} X^{(1)}_{ijk}$$

$$H^{(1,2)} = \sum_i \sum_j \left[\sum_k h_{ijk} X^{(1)}_{ijk} \right] X^{(2)}_{ij}$$

$$H^{(1,3)} = \sum_i \sum_k \left[\sum_j h_{ijk} X^{(1)}_{ijk} \right] X^{(3)}_{ik}$$

$$H^{(1,4)} = \sum_j \sum_k \left[\sum_i h_{ijk} X^{(1)}_{ijk} \right] X^{(4)}_{jk}$$

$$H^{(1,5)} = \sum_i \left[\sum_j \sum_k h_{ijk} X^{(1)}_{ijk} \right] X^{(5)}_{i}$$

$$H^{(1,6)} = \sum_j \left[\sum_i \sum_k h_{ijk} X^{(1)}_{ijk} \right] X^{(6)}_{j}$$

$$H^{(1,7)} = \sum_k \left[\sum_i \sum_j h_{ijk} X^{(1)}_{ijk} \right] X^{(7)}_{k}$$

$$H^{(2,2)} = \sum_i \sum_j \left[\sum_k h_{ijk} \right] X^{(2)}_{ij} X^{(2)}_{ij}$$

$$H^{(2,3)} = \sum_k \sum_i \left[\sum_j h_{ijk} X^{(2)}_{ij} \right] X^{(3)}_{ik}$$

$$H^{(2,4)} = \sum_k \sum_j \left[\sum_i h_{ijk} X^{(2)}_{ij} \right] X^{(4)}_{jk}$$

$$H^{(2,5)} = \sum_i \left[\sum_j \left(\sum_k h_{ijk} \right) X^{(2)}_{ij} \right] X^{(5)}_{i}$$

$$H^{(2,6)} = \sum_j \left[\sum_i \left(\sum_k h_{ijk} \right) X^{(2)}_{ij} \right] X^{(6)}_{j}$$

$$H^{(2,7)} = \sum_k \left[\sum_i \sum_j h_{ijk} X^{(2)}_{ij} \right] X^{(7)}_{k}$$

$$H^{(3,3)} = \sum_k \sum_i \left[\sum_j h_{ijk} \right] X^{(3)}_{ik} X^{(3)}_{ik}$$

$$H^{(3,4)} = \sum_k \sum_j \left[\sum_i h_{ijk} X^{(3)}_{ik} \right] X^{(4)}_{jk}$$

$$H^{(3,5)} = \sum_i \left[\sum_k \left(\sum_j h_{ijk} \right) X^{(3)}_{ik} \right] X^{(5)}_{i}$$

$$H^{(3,6)} = \sum_j \left[\sum_k \sum_i h_{ijk} X^{(3)}_{ik} \right] X^{(6)}_{j}$$

$$H^{(3,7)} = \sum_k \left[\sum_i \left(\sum_j h_{ijk} \right) X^{(3)}_{ik} \right] X^{(7)}_{k}$$

$$H^{(4,4)} = \sum_k \sum_j \left[\sum_i h_{ijk} \right] X^{(4)}_{jk} X^{(4)}_{jk}$$

$$H^{(4,5)} = \sum_i \left[\sum_k \sum_j h_{ijk} X^{(4)}_{jk} \right] X^{(5)}_{i}$$

$$H^{(4,6)} = \sum_j \left[\sum_k \left(\sum_i h_{ijk} \right) X^{(4)}_{jk} \right] X^{(6)}_{j}$$

$$H^{(4,7)} = \sum_k \left[\sum_j \left(\sum_i h_{ijk} \right) X^{(4)}_{jk} \right] X^{(7)}_{k}$$

$$H^{(5,5)} = \sum_i \left[\sum_k \sum_j h_{ijk} \right] X^{(5)}_{i} X^{(5)}_{i}$$

$$H^{(5,6)} = \sum_j \left[\sum_i \left(\sum_k h_{ijk} \right) X^{(5)}_{i} \right] X^{(6)}_{j}$$

$$H^{(5,7)} = \sum_i \left[\sum_k \left(\sum_j h_{ijk} \right) X^{(7)}_{k} \right] X^{(5)}_{i}$$

$$H^{(6,6)} = \sum_j \left[\sum_k \sum_i h_{ijk} \right] X^{(6)}_{j} X^{(6)}_{j}$$

$$H^{(6,7)} = \sum_j \left[\sum_k \left(\sum_i h_{ijk} \right) X^{(7)}_{k} \right] X^{(6)}_{j}$$

$$H^{(7,7)} = \sum_k \left[\sum_i \sum_j h_{ijk} \right] X^{(7)}_{k} X^{(7)}_{k}$$

$$H^{(i,j)} = H^{(j,i)} \quad \text{if } i > j$$

2 Multivariate generalized linear model

2.1 Setup

Model:

$$\mathbf{Y}_n \sim \mathcal{F}(\mathbf{V}_n) \qquad \text{where} \quad \mathbf{V}_n = \mathbf{X}_n \beta \tag{17}$$

where $\mathbf{Y}_n$ and $\mathbf{V}_n$ are vectors, $\mathcal{F}$ is some probability distribution, $\mathbf{X}_n$ is a $[J \times L]$ vector of observables, and β is a $[L \times 1]$ vector of parameters to estimate.

2.2 Multinomial Logit Model

Notations:

- $\mathbf{Y}_n$ is a $[J \times 1]$ vector such that $Y_{nj} \in \{0, 1\}$ for all j
- $M_n = \sum_j \mathbf{Y}_{nj}$ is the number of Multinomial trials
- $\mathbf{X}_n$ is a $[J \times L]$ matrix of covariates
- β is a $[L \times 1]$ matrix of parameters
- $\mathbf{V}_n$ is a $[J \times 1]$ vector of utilities that defines outcome probabilities $\mathbf{p}_n$
- $\mathbf{p}_n$ is a $[J \times 1]$ vector of probabilities such that $\sum_j p_{nj} = 1$ and $0 \leq p_{nj} \leq 1$ for all j
- $\mathbf{y}$ is a $[NJ \times 1]$ vector that stacks up all values y_{nj}
- $\mathbf{p}$ is a $[NJ \times 1]$ vector that stacks up all values p_{nj}
- $\mathbf{logp}$ is a $[NJ \times 1]$ vector that stacks up all values $\log(p_{nj})$
- $\mathbf{M}$ is a $[N \times 1]$ vector that collects the values M_n
- $\tilde{\mathbf{M}}$ is a $[NJ \times 1]$ vector that repeats the values M_n, such that $\tilde{M}_{nj} = M_n$ for all j
- $\mathbf{A}$ is a $[N \times L]$ matrix such that $A_{nl} = \sum_j p_{nj} X_{njl}$

Model:

$$\begin{aligned} \mathbf{Y}_n &\sim Multinomial(M_n, \mathbf{p}_n) \\ p_{nj} &= \exp(V_{nj}) \Big/ \sum_k \exp(V_{nk}) \text{ for all } j \\ \mathbf{V}_n &= \mathbf{X}_n \beta \end{aligned} \tag{18}$$

Log-likelihood:

$$\begin{aligned} LL &= \sum_{j=1}^{J} y_{nj} \log(p_{nj}) + C = \boxed{\mathbf{y}'\mathbf{logp} + C} \\ \text{where} \quad C &= \sum_n \left[\log(M_n!) - \sum_{j=1}^{J} \log(y_{nj}!) \right] \end{aligned} \tag{19}$$

Gradient of log-likelihood:

$$\frac{\partial LL}{\partial \beta_l} = \sum_n \sum_{j=1}^{J} X_{njl}(y_{nj} - M_n p_{nj}) \implies \nabla LL(\beta) = \boxed{\mathbf{X}'\left(\mathbf{y} - \tilde{\mathbf{M}} \odot \mathbf{p}\right)} \tag{20}$$

where $\odot$ represents the Hadamard product (term-by-term multiplication).

Hessian of log-likelihood:

$$\begin{aligned}
\frac{\partial^2 LL}{\partial \beta_l \partial \beta_{l'}} &= -\sum_{n,j} \tilde{M}_{nj} p_{nj} X_{njl} X_{njl'} + \sum_n M_n A_{nl} A_{nl'} \\
\text{where} \quad A_{nl} &= \sum_j p_{nj} X_{njl} \\
\implies \mathbf{H}(\beta) &= \boxed{\mathbf{X}' \left[\mathbf{X} \odot (\mathbf{h} \cdot \mathbb{1}_L)\right] + \mathbf{A}' \left[\mathbf{A} \odot (\mathbf{M} \cdot \mathbb{1}_L)\right]} \\
\text{where } \mathbf{h} &= -\tilde{\mathbf{M}} \odot \mathbf{p}
\end{aligned} \tag{21}$$

2.3 Estimation by Maximum Likelihood

Computational trick:

$$\begin{aligned}
\log(p_{nj}) &= V_{nj} - \log\left(\sum_k \exp(V_{nk})\right) \\
&= V_{nj} - \bar{V}_n - \log\left(\sum_k \exp(V_{nk} - \bar{V}_n)\right) \\
\text{where} \quad \bar{V}_n &= \max_j \; V_{nj}
\end{aligned} \tag{22}$$

This trick avoids overflow issues that arise when computing the exponential of large values.