Week 4 Notes: GLM Theory

This week, we will learn about some of the basic theory underlying GLMs, as well as how these models tend to be fit in practice. While somewhat abstract, it's worth noting that the topics presented here have highly pragmatic use cases, from constructing hypothesis tests to designing inference algorithms, and I think it's important that one see these things at least once in their life.

1 The Likelihood of a GLM

GLMs are fit in `R` using *likelihood based inference.* The likelihood function for a GLM, given data $\mathcal{D} = \{(Y_i, X_i) : i = 1, \dots, N\}$ is given by

$$L(\beta, \phi) = \prod_{i=1}^{N} \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(Y_i; \phi/\omega_i)\right\},$$

where we define $\theta_i \equiv (b')^{-1}(\mu_i)$ and $\mu_i \equiv g^{-1}(X_i^\top \beta)$. We can then derive the *score function* $s(\beta, \phi) = \frac{\partial}{\partial\beta} \log L(\beta, \phi)$ as

$$s(\beta, \phi) = \sum_{i=1}^{N} \frac{\partial}{\partial\beta} \frac{Y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(Y_i; \phi/\omega_i).$$

Again, we will write $\frac{\partial}{\partial\beta}F(\beta)$ for the gradient of F and $\frac{\partial^2}{\partial\beta\partial\beta^\top}F(\beta)$ for the Hessian matrix.

Exercise 1: Deriving the Score

Using the chain rule $\frac{\partial}{\partial\beta} = \frac{\partial}{\partial\theta} \times \frac{\partial\theta}{\partial\mu} \times \frac{\partial\mu}{\partial\beta}$, show that

$$s(\beta, \phi) = \sum_{i=1}^{N} \frac{\omega_i(Y_i - \mu_i)X_i}{\phi V(\mu_i)g'(\mu_i)}.$$

Show also that, for the canonical link, we have $g'(\mu_i) = V(\mu_i)^{-1}$ so that this reduces to

$$s(\beta, \phi) = \sum_{i=1}^{N} \frac{\omega_i (Y_i - \mu_i) X_i}{\phi}.$$

Hint: recall that $\frac{d}{dx} g^{-1}(x) = \frac{1}{g'\{g^{-1}(x)\}}$.

Note for posterity that the MLE is justified, in large part, because it is the sample solution $s(\widehat{\beta}, \phi) = 0$ to the population-level estimating equation $\int s(\beta, \phi)\, f_0(\mathbf{y} \mid \mathbf{X}) d\mathbf{y} = 0$ (here, $\mathbf{X}$ denotes the design matrix and f_0 denotes the true conditional density of $\mathbf{Y} = (Y_1, \ldots, Y_N)$. Also note that this estimating equation holds *even when the model is misspecified!* We only need the mean structure (not the exponential dispersion family) for the population-level estimating equation to hold, which suggests that the MLE $\widehat{\beta}$ might still be a good estimator even when the model is misspecified. More on this when we study M-estimation.

Exercise 2: Deriving the Fisher Information

We define the *Fisher Information* to be

$$\mathcal{I}(\beta, \phi) = -\mathbb{E}\left\{ \frac{\partial^2}{\partial \beta \partial \beta^\top} \log L(\beta, \phi) \mid \beta, \phi \right\}.$$

The Fisher information plays an important role in inference for GLMs. The "observed" Fisher information is also used,

$$\mathcal{J}(\beta, \phi) = -\frac{\partial^2}{\partial \beta \partial \beta^\top} \log L(\beta, \phi).$$

In addition to being easier to evaluate, using $\mathcal{J}$ has been argued to be the right-thing-to-do™. In any case, show that

$$\langle \mathcal{J}(\beta, \phi) \rangle_{jk} = \frac{1}{\phi} \sum_{i=1}^{N} X_{ij} X_{ik} \left\{ \frac{\omega_i}{V(\mu_i) g'(\mu_i)^2} - \frac{\omega_i (Y_i - \mu_i)}{g'(\mu_i)} \left(\frac{\partial}{\partial \mu_i} \frac{1}{V(\mu_i) g'(\mu_i)} \right) \right\}$$

and

$$\langle \mathcal{I}(\beta, \phi) \rangle_{jk} = \frac{1}{\phi} \sum_{i=1}^{N} X_{ij} X_{ik} \frac{\omega_i}{V(\mu_i) g'(\mu_i)^2}$$

where $\langle A \rangle_{ij}$ denotes the $(i, j)^{\text{th}}$ element of the matrix A. Show also that $\mathcal{I}(\beta, \phi) = \mathcal{J}(\beta, \phi)$ when the canonical link is used.

From the above exercise, notice that the Fisher information has the familiar form

$$\mathcal{I}^{-1} = \phi(\mathbf{X}^\top D \mathbf{X})^{-1}$$

where D is a diagonal matrix with entries $\omega_i / \{V(\mu_i) g(\mu_i)^2\}$. Similarly, we can write $\mathcal{J}^{-1} = \phi(\mathbf{X}^\top \widetilde{D} \mathbf{X})^{-1}$ for some diagonal matrix $\widetilde{D}$. Compare this with the linear model, which has inverse Fisher information $\mathcal{J}^{-1} = \sigma^2 (X^\top X)^{-1}$.