Lecture 27 — Poisson regression

27.1 The Poisson log-linear model

Example 27.1. Neurons in the central nervous system transmit signals via a series of action potentials, or "spikes". The spiking of a single neuron may be measured by a microelectrode, and its sequence of spikes over time is called a spike train. A simple and commonly-used statistical model for a spike train is an inhomogeneous Poisson point process, which has the following property: For n time windows of length Δ, letting Y_i denote the number of spikes generated by the neuron in the i^{th} time window, the random variables $Y_1, \ldots, Y_n$ are independent and distributed as $Y_i \sim \text{Poisson}(\lambda_i \Delta)$, where the parameter λ_i controls the spiking rate in the i^{th} time window. For simplicity, we will assume $\Delta = 1$.

The spiking rate λ_i of a neuron may be influenced by external sensory stimuli present in this i^{th} window of time, for example the intensity and pattern of light visible to the eye or the texture of an object presented to the touch. To understand the effects of these sensory stimuli on the spiking rate of a particular neuron, we may perform an experiment that applies different stimuli in different windows of time and records the neural response. Encoding the stimuli applied in the i^{th} window of time by a set of p covariates $x_{i1}, \ldots, x_{ip}$, a simple model for the Poisson rate parameter λ_i is given by

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \tag{27.1}$$

or equivalently,

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}.$$

Together with the distributional assumption $Y_i \sim \text{Poisson}(\lambda_i)$, this is called the **Poisson log-linear model**, or the Poisson regression model. It is a special case of what is known in neuroscience as the linear-nonlinear Poisson cascade model.

More generally, the Poisson log-linear model is a model for n responses $Y_1, \ldots, Y_n$ that take integer count values. Each Y_i is modeled as an independent $\text{Poisson}(\lambda_i)$ random variable, where $\log \lambda_i$ is a linear combination of the covariates corresponding to the i^{th} observation. As in the cases of linear and logistic regression, we treat the covariates as fixed constants, and the model parameters to be inferred are the regression coefficients $\beta = (\beta_0, \ldots, \beta_p)$.

27.2 Statistical inference

We will describe the procedure for maximum-likelihood estimation of the regression coefficients and Fisher-information based estimation of their standard errors, and discuss some issues concerning model misspecification and robust standard error estimates.

Since $Y_1, \ldots, Y_n$ are independent Poisson random variables, the likelihood function is given by

$$\text{lik}(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{\lambda_i^{Y_i} e^{-\lambda_i}}{Y_i!}$$

where λ_i is defined in terms of $\beta_0, \ldots, \beta_p$ and the covariates $x_{i1}, \ldots, x_{ip}$ via equation (27.1). Setting $x_{i0} \equiv 1$ for all i, the log-likelihood is then

$$\begin{aligned} l(\beta_0, \ldots, \beta_p) &= \sum_{i=1}^{n} Y_i \log \lambda_i - \lambda_i - \log Y_i! \\ &= \sum_{i=1}^{n} Y_i \left(\sum_{j=0}^{p} \beta_j x_{ij} \right) - e^{\sum_{j=0}^{p} \beta_j x_{ij}} - \log Y_i! \end{aligned}$$

and the MLEs are the solutions to the system of score equations, for $m = 0, \ldots, p$,

$$0 = \frac{\partial l}{\partial \beta_m} = \sum_{i=1}^{n} x_{im}(Y_i - e^{\sum_{j=0}^{p} \beta_j x_{ij}}).$$

These equations may be solved numerically using the Newton-Raphson method.

The Fisher information matrix $I_{\mathbf{Y}}(\beta) = -\mathbb{E}_\beta[\nabla^2 l(\beta)]$ may be obtained by computing the second-order partial derivatives of l:

$$\frac{\partial^2 l}{\partial \beta_m \partial \beta_l} = -\sum_{i=1}^{n} x_{im} x_{il} e^{\sum_{j=0}^{p} \beta_j x_{ij}}.$$

Writing $X_j = (x_{1j}, \ldots, x_{nj})$ as the jth column of the covariate matrix X and defining the diagonal matrix

$$W = W(\beta) := \text{diag}\left(e^{\sum_{j=0}^{p} \beta_j x_{1j}}, \ldots, e^{\sum_{j=0}^{p} \beta_j x_{nj}} \right),$$

the above may be written as $\frac{\partial^2 l}{\partial \beta_m \partial \beta_l} = -X_m^T W X_l$, so $\nabla^2 l(\beta) = -X^T W X$ and $I_{\mathbf{Y}}(\beta) = X^T W X$. For large n, if the Poisson log-linear model is correct, then the MLE vector $\hat{\beta}$ is approximately distributed as $\mathcal{N}(\beta, (X^T W X)^{-1})$. We may then estimate the standard error of $\hat{\beta}_j$ by

$$\hat{\text{se}}_j = \sqrt{((X^T \hat{W} X)^{-1})_{jj}},$$

where $\hat{W} = W(\hat{\beta})$ is the plugin estimate for W. These formulas are the same as for the case of logistic regression in Lecture 26, except with a different form of the diagonal matrix W.

The modeling assumption of a Poisson distribution for Y_i is rather restrictive, as it implies that the variance of Y_i must be equal to its mean. This is rarely true in practice, and it is frequently the case that the observed variance of Y_i is larger than its mean—this problem is known as **overdispersion**. Nonetheless, the Poisson regression model is oftentimes used in overdispersed settings: As long as $Y_1, \ldots, Y_n$ are independent and

$$\log \mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

for each i (so the model for the means of the Y_i's is correct), then it may be shown that the MLE $\hat{\beta}$ in the Poisson regression model is unbiased for β, even if the distribution of Y_i is not Poisson and the variance of Y_i exceeds its mean. The above standard error estimate $\hat{\text{se}}_j$ and the associated confidence interval for β_j, though, would not correct in the overdispersed setting. One may use instead the robust sandwich estimate of the covariance of $\hat{\beta}$, given by

$$(X^T \hat{W} X)^{-1} (X^T \tilde{W} X)(X^T \hat{W} X)^{-1}$$

where

$$\tilde{W} = \text{diag}((Y_1 - \hat{\lambda}_1)^2, \ldots, (Y_n - \hat{\lambda}_n)^2)$$

and $\hat{\lambda}_i = e^{\sum_{j=0}^{p} \hat{\beta}_j x_{ij}}$ is the fitted value of λ for the i^{th} observation. Alternatively, one may use the pairs bootstrap procedure as described in Lecture 26.

Remark 27.2. The linear model, logistic regression model, and Poisson regression model are all examples of the **generalized linear model (GLM)**. In a generalized linear model, $Y_1, \ldots, Y_n$ are modeled as independent observations with distributions $Y_i \sim f(y|\theta_i)$ for some one-parameter family $f(y|\theta)$. The parameter θ_i is modeled as

$$g(\theta_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

for some one-to-one transformation $g : \mathbb{R} \to \mathbb{R}$ called the **link function**, where $x_{i1}, \ldots, x_{ip}$ are covariates corresponding to Y_i. In the linear model considered in Lecture 25, the parameter was $\theta \equiv \mu$ where $f(y|\mu)$ was the PDF of the $\mathcal{N}(\mu, \sigma_0^2)$ distribution (for a known variance σ_0^2), and $g(\mu) = \mu$. In logistic regression, the parameter was $\theta \equiv p$ where $f(y|p)$ was the PMF of the Bernoulli(p) distribution, and $g(p) = \log \frac{p}{1-p}$. In Poisson regression, the parameter was $\theta \equiv \lambda$ where $f(y|\lambda)$ was the PMF of the Poisson(λ) distribution, and $g(\lambda) = \log \lambda$.

The choice of the link function g is an important modeling decision, as it determines which transform of the model parameter should be modeled as linear in the observed covariates. In each of the three examples discussed, we used what is called the **natural link**, which is motivated by considering a change-of-variable for the parameter, $\theta \mapsto \eta(\theta)$, so that the PDF/PMF $f(y|\eta)$ in terms of the new parameter η has the form

$$f(y|\eta) = e^{\eta y - A(\eta)} h(y)$$

for some functions A and h. For example, the Bernoulli PMF is

$$f(y) = p^y (1-p)^{1-y} = (1-p) \left(\frac{p}{1-p} \right)^y = e^{\left(\log \frac{p}{1-p}\right) y + \log(1-p)},$$

so we may set $\eta = \log \frac{p}{1-p}$, $A(\eta) = -\log(1-p) = \log(1 + e^\eta)$, and $h(y) = 1$. This is called the **exponential family** form of the PDF/PMF, and η is called the **natural parameter**. In each example, the natural link simply sets $g(\theta) = \eta(\theta)$ (or equivalently, $g(\theta) = c\eta(\theta)$ for a constant c).

Use of the natural link leads to some nice mathematical properties for likelihood-based inference—for instance, since η is modeled as linear in β, the second-order partial derivatives of

$$\log f(Y|\eta) = \eta Y - A(\eta) + \log h(Y)$$

with respect to β do not depend on Y, so the Fisher information is always given by $-\nabla^2 l(\beta)$ without needing to take an expectation. (We sometimes say in this case that the "observed and expected Fisher information matrices" are the same.) On the other hand, from the modeling perspective, there is usually no intrinsic reason to believe that the natural link $g(\theta) = \eta(\theta)$ is the correct transformation of θ that is well-modeled as a linear combination of the covariates, and other link functions are also commonly used, especially if they lead to a better fit for the data.