Generalized Linear Models (Part III)

Rebecca C. Steorts (adaptation from notes of Yue Jiang)

March 3, 2025

Poisson Data

Suppose we model our data as coming from a Poisson distribution with parameter λ:

$$f_Y(y) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

Question: How might you estimate λ given our observed data?

Review: Maximum Likelihood Estimation

Assuming the observations $Y_1, Y_2, \ldots, Y_n$ are i.i.d., the likelihood is

$$L(\lambda \mid Y) = \prod_{i=1}^{n} f(Y_i \mid \lambda).$$

Recall that the likelihood is the probability of the observed data given λ. (Do not confuse $f(Y_i|\lambda)$ with $f(\lambda|Y_i)$.)

Question: If $Y_i \sim \text{Pois}(\lambda)$ (iid), what is the MLE for λ?

MLE for the Poisson Distribution

The likelihood function is:

$$L(\lambda \mid Y) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}.$$

Taking logs,

$$\begin{aligned} \log L(\lambda \mid Y) &= \sum_{i=1}^{n} \Big(y_i \log \lambda - \lambda - \log(y_i!) \Big) && (1) \\ &= \left(\log \lambda \sum_{i=1}^{n} y_i \right) - n\lambda - \sum_{i=1}^{n} \log(y_i!). && (2) \end{aligned}$$

MLE for the Poisson Distribution

Setting the derivative (score function) to zero:

$$\frac{\partial}{\partial\lambda} \log L(\lambda \mid Y) = \frac{1}{\lambda}\sum_{i=1}^{n} y_i - n = 0,$$

which gives

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Verification of Maximum

Check the second derivative:

$$\frac{\partial^2}{\partial\lambda^2} \log L(\lambda \mid Y) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} y_i - n < 0.$$

Thus, $\hat{\lambda}$ is indeed a maximum.

Poisson Regression

We extend the Poisson model to incorporate covariates using a generalized linear model:

$$\log\big(\underbrace{E(Y \mid X)}_{\lambda}\big) = X^T\beta,$$

where we assume the outcome is Poisson and the canonical link is the logarithm.

Question: Can we differentiate the log-likelihood, set it equal to zero, and solve for the MLEs of $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ as before?

Poisson Regression Log-Likelihood

The log-likelihood is:

$$\log L = \sum_{i=1}^{n} \Big(y_i \log \lambda - \lambda - \log(y_i!) \Big) \tag{3}$$

$$= \sum_{i=1}^{n} \Big(y_i X_i^T \beta - e^{X_i^T \beta} - \log(y_i!) \Big). \tag{4}$$

In general, setting the score equations

$$\frac{\partial \log L}{\partial \beta_j} = 0,$$

leads to what we call "transcendental equations," that typically have no closed-form solution.

Newton-Raphson Algorithm (1D)

Newton-Raphson for root finding is based on a second-order Taylor approximation:

1. Start with an initial guess $\theta^{(0)}$.
2. Iterate:
$$\theta^{(t+1)} = \theta^{(t)} - \frac{f'(\theta^{(t)})}{f''(\theta^{(t)})}.$$
3. Stop when a convergence criterion is satisfied.

There are some necessary conditions for convergence, however, this is beyond the scope of this class. Many likelihood functions you are likely to encounter (e.g., GLMs with canonical link) will in fact converge from any starting value.

Newton-Raphson in Higher Dimensions

Define the score vector and Hessian for $logL(\theta \mid X)$ with parameter vector $\theta = (\theta_1, \ldots, \theta_p)$ as follows:

$$\nabla \log L = \begin{pmatrix} \frac{\partial \log L}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log L}{\partial \theta_p} \end{pmatrix}, \quad \nabla^2 \log L = \begin{pmatrix} \frac{\partial^2 \log L}{\partial \theta_1^2} & \cdots & \frac{\partial^2 \log L}{\partial \theta_1 \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L}{\partial \theta_p \theta_1} & \cdots & \frac{\partial^2 \log L}{\partial \theta_p^2} \end{pmatrix}.$$

Then the update is:

$$\theta^{(t+1)} = \theta^{(t)} - \left[\nabla^2 \log L(\theta^{(t)} \mid X)\right]^{-1} \nabla \log L(\theta^{(t)} \mid X).$$

Newton-Raphson in Higher Dimensions

We can modify the Newton-Raphson algorithms for higher dimensions as follows:

1. Start with an initial guess $\theta^{(0)}$.
2. Iterate

$$\theta^{(t+1)} = \theta^{(t)} - \left[\nabla^2 \log L(\theta^{(t)} \mid X)\right]^{-1} \nabla \log L(\theta^{(t)} \mid X).$$

3. Stop when convergence criterion is satisfied.

Newton-Raphson for Poisson Regression

For Poisson regression:

$$\log L = \sum_{i=1}^{n} \Big(y_i X_i \beta - e^{X_i \beta} - \log(y_i!) \Big),$$

What are the score vector and Hessian corresponding to the log-likelihood? What would the Newton-Raphson steps be?

Newton-Raphson for Poisson Regression

For Poisson regression:

$$\log L = \sum_{i=1}^{n} \Big(y_i X_i \beta - e^{X_i \beta} - \log(y_i!) \Big),$$

the score vector is:

$$\nabla \log L = \sum_{i=1}^{n} \Big(y_i - e^{X_i \beta} \Big) X_i^T,$$

and the Hessian is:

$$\nabla^2 \log L = -\sum_{i=1}^{n} e^{X_i \beta} X_i X_i^T.$$

Thus, the Newton-Raphson update becomes:

$$\beta^{(t+1)} = \beta^{(t)} - \Big[-\sum_{i=1}^{n} e^{X_i^T \beta} X_i X_i^T \Big]^{-1} \sum_{i=1}^{n} \Big(y_i - e^{X_i^T \beta} \Big) X_i^T.$$

Exercise 1

Consider the linear regression model under the normality assumption (and constant variance). Is this a GLM? If so, identify the three components needed. If not, explain why not.

Exercise 2

Suppose we're trying to model the number of cancer cases per month Y in a city, conditionally on various demographic and exposure factors. Consider the log-linear regression model $\log(E[Y \mid X]) = X\beta$, where Y takes on a Poisson distribution with parameter λ. Is this a GLM? If so, identify the three components needed (including specifics regarding the exponential family) and specifically identify whether the link function is canonical. If not, explain why not.

Exercise 3

Suppose we're trying to model the waiting time until the next bus arrives Y, conditionally on weather conditions and traffic. Consider the log-linear regression model $\log(E[Y \mid X]) = X\beta$, where Y takes on an Exponential distribution with parameter λ. Is this a GLM? If so, identify the three components needed (including specifics regarding the exponential family) and specifically identify whether the link function is canonical. If not, explain why not.

Exercise 4

Derive the score and Hessian functions of the log-likelihood for a logistic regression model (i.e., binary regression under canonical link).

Solution: Exercise 1

Consider the linear regression model under the normality assumption (and constant variance). Is this a GLM? If so, identify the three components needed. If not, explain why not.

Solution: Yes. Under the normality and constant variance assumptions, the distribution of $Y \mid X$ is $N(X\beta, \sigma^2)$. The conditional expectation $E(Y \mid X)$ is linked through the linear predictor $X\beta$ through the identity function.

Solution: Exercise 2

Suppose we're trying to model the number of cancer cases per month Y in a city, conditionally on various demographic and exposure factors. Consider the log-linear regression model $\log(E[Y \mid X]) = X\beta$, where Y takes on a Poisson distribution with parameter λ. Is this a GLM? If so, identify the three components needed (including specifics regarding the exponential family) and specifically identify whether the link function is canonical. If not, explain why not.

Yes, the distribution of $Y \mid X \sim Pois(\lambda)$. The conditional expectation $\lambda = E(Y \mid X)$ is linked through the linear predictor $X\beta$ through the log function. This is the canonical link. The relevant functions for the Poisson distribution in canonical form are as follows

Solution: Exercise 2

$$h(y) = \frac{1}{y!} I(y \in \{0, 1, 2, \cdots\}).$$

$$\eta(\lambda) = \log(\lambda) \implies \lambda = e^{\lambda}.$$

$$T(y) = y$$

$$A(\eta) = \lambda$$

Solution: Exercise 3

Suppose we're trying to model the waiting time until the next bus arrives Y, conditionally on weather conditions and traffic. Consider the log-linear regression model $\log(E[Y \mid X]) = X\beta$, where Y takes on an Exponential distribution with parameter λ. Is this a GLM? If so, identify the three components needed (including specifics regarding the exponential family) and specifically identify whether the link function is canonical. If not, explain why not.

Yes, the distribution of Y is $Exp(\lambda)$. The conditional expectation $\lambda = E(Y \mid X)$ is linked to the linear predictor $X\beta$ through the log function. However, this is not the canonical link, since the relevant parameters can be shown to be as follows:

Solution: Exercise 3

$$h(y) = I(y > 0)$$

$$-\lambda = \eta \implies \lambda = -\eta$$

$$T(y) = y$$

$$A(\eta) = -\log(-\eta)$$

Exercise 4

Derive the score and Hessian functions of the log-likelihood for a logistic regression model (i.e., binary regression under canonical link).

Solution Exercise 4

The log-likelihood function for logistic regression is

$$
\begin{aligned}
\mathcal{L}(p) &= \prod_{i=1}^{n} f(y_i) \\
&= \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}, \\
\log \mathcal{L}(p) &= \sum_{i=1}^{n} \Big[y_i \log(p) + \log(1-p) - y_i \log(1-p) \Big] \\
&= \sum_{i=1}^{n} \Big[y_i \log\left(\frac{p}{1-p} \right) + \log(1-p) \Big].
\end{aligned}
$$

Solution Exercise 4

In logistic regression we have

$$\log\left(\frac{p}{1-p}\right) = \mathbf{X}\boldsymbol{\beta},$$

so the log-likelihood becomes

$$\log \mathcal{L}(p) = \sum_{i=1}^{n}\Big[y_i\,\mathbf{X}_i\boldsymbol{\beta} - \log\big(1+\exp\big(\mathbf{X}_i\boldsymbol{\beta}\big)\big)\Big].$$

Differentiating with respect to $\boldsymbol{\beta}$ gives the score function:

$$\nabla_{\boldsymbol{\beta}} \log \mathcal{L}(p) = \sum_{i=1}^{n}\left[y_i\,\mathbf{X}_i - \frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\,\mathbf{X}_i\right],$$

and the Hessian is

$$\nabla^2_{\boldsymbol{\beta}} \log \mathcal{L}(p) = -\sum_{i=1}^{n}\left[\frac{1}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\cdot\frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})}\,\mathbf{X}_i\,\mathbf{X}_i^T\right].$$