Unit 5: Generalized linear models: Special cases

Eugene Katsevich

November 10, 2021

Unit 4 developed a general theory for GLMs. In Unit 5, we specialize this theory to several important cases, including logistic regression and Poisson regression.

1 Logistic regression

1.1 Model definition and interpretation

Model definition. Recall from Unit 4 that the logistic regression model is

$$m_i y_i \overset{\text{ind}}{\sim} \text{Bin}(m_i, \pi_i); \quad \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = \boldsymbol{x}_{i*}^T \boldsymbol{\beta}. \tag{1}$$

Here we use the canonical logit link function, although other link functions are possible. The interpretation of the parameter β_j is that a unit increase in x_j—other predictors held constant—is associated with an (additive) increase of β_j on the log-odds scale or a multiplicative increase of e^{β_j} on the odds scale. Note that logistic regression data come in two formats: *ungrouped* and *grouped*. For ungrouped data, we have $m_1 = \cdots = m_n = 1$, so $y_i \in \{0, 1\}$ are Bernoulli random variables. For grouped data, we can have several independent Bernoulli observations per predictor $\boldsymbol{x}_{i*}$, which give rise to binomial proportions $y_i \in [0, 1]$. This happens most often when all the predictors are discrete. You can always convert grouped data into ungrouped data, but not necessarily vice versa. We'll discuss below that the grouped and ungrouped formulations of logistic regression have the same MLE and standard errors but different deviances.

Generative model equivalent. Consider the following generative model for $(\boldsymbol{x}, y) \in \mathbb{R}^{p-1} \times \{0, 1\}$:

$$y \sim \text{Ber}(\pi); \quad \boldsymbol{x}|y \sim \begin{cases} N(\boldsymbol{\mu}_0, \boldsymbol{V}) & \text{if } y = 0 \\ N(\boldsymbol{\mu}_1, \boldsymbol{V}) & \text{if } y = 1 \end{cases}. \tag{2}$$

Then, we can derive that $y|\boldsymbol{x}$ follows a logistic regression model (called a *discriminative* model because it conditions on $\boldsymbol{x}$). Indeed,

$$\begin{aligned} \text{logit}(p(y=1|\boldsymbol{x})) &= \log \frac{p(y=1)p(\boldsymbol{x}|y=1)}{p(y=0)p(\boldsymbol{x}|y=0)} \\ &= \log \frac{\pi \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)}{(1-\pi)\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\right)} \\ &= \beta_0 + \boldsymbol{x}^T \boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \\ &\equiv \beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}_{\text{-}0}. \end{aligned} \tag{3}$$

This is another natural route to motivating the logistic regression model.

Special case: 2×2 contingency table. Suppose that $x \in \{0, 1\}$, and consider the logistic regression model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$. For example, suppose that $x \in \{0, 1\}$ encodes treatment (1) and control (0) in a clinical trial, and $y_i \in \{0, 1\}$ encodes success (1) and failure (0). We make n observations of (x_i, y_i) in this ungrouped setup. The parameter e^{β_1} can be interpreted as the *odds ratio*:

$$e^{\beta_1} = \frac{\mathbb{P}[y = 1|x = 1]/\mathbb{P}[y = 0|x = 1]}{\mathbb{P}[y = 1|x = 0]/\mathbb{P}[y = 0|x = 0]}. \tag{4}$$

This parameter is the multiple by which the odds of success increase when going from control to treatment. We can summarize such data via the 2×2 *contingency table* (Table 1). A grouped version of this data would be $\{(x_1, y_1) = (0, 7/24), (x_2, y_2) = (1, 9/21)\}$. The null hypothesis $H_0 : \beta_1 = 0 \iff H_0 : e^{\beta_1} = 1$ states that the success probability in both rows of the table is the same.

	Success	Failure	Total
Treatment	9	12	21
Control	7	17	24
Total	16	29	45

Table 1: An example of a 2×2 contingency table.

Logistic regression with case-control studies. In a prospective study (e.g. a clinical trial), we assign treatment or control (i.e., x) to individuals, and then observe a binary outcome (i.e., y). Sometimes, the outcome y takes a long time to measure or has highly imbalanced distribution in the population (e.g. the development of lung cancer). In this case, an appealing study design is the *retrospective study*, where individuals are sampled based on their *response values* (e.g. presence of lung cancer) rather than their treatment/exposure status (e.g. smoking). It turns out that a logistic regression model is appropriate for such retrospective study designs as well. Indeed, suppose that $y|\boldsymbol{x}$ follows a logistic regression model. Let's try to figure out the distribution of $y|\boldsymbol{x}$ in the retrospectively gathered sample. Letting $z \in \{0, 1\}$ denote the indicator that an observation is sampled, define $\rho_1 \equiv \mathbb{P}[z = 1|y = 1]$ and $\rho_0 \equiv \mathbb{P}[z = 1|y = 0]$, and assume that $\mathbb{P}[z = 1, y, \boldsymbol{x}] = \mathbb{P}[z = 1|y]$. The latter assumption states that the predictors $\boldsymbol{x}$ were not used in the retrospective sampling process. Then,

$$\text{logit}(\mathbb{P}[y = 1|z = 1, \boldsymbol{x}]) = \log \frac{\rho_1 \mathbb{P}[y = 1|\boldsymbol{x}]}{\rho_0 \mathbb{P}[y = 0|\boldsymbol{x}]} = \log \frac{\rho_1}{\rho_0} + \text{logit}(\mathbb{P}[y = 1|\boldsymbol{x}]) = \left(\log \frac{\rho_1}{\rho_0} + \beta_0\right) + \boldsymbol{x}^T \boldsymbol{\beta}_{\text{-}0}.$$

Thus, conditioning on retrospective sampling changes only the intercept term, but preserves the coefficients of $\boldsymbol{x}$. Therefore, we can carry out inference for $\boldsymbol{\beta}_{\text{-}0}$ in the same way regardless of whether the study design is prospective or retrospective.

1.2 Estimation and inference

Score and Fisher information. We recall from Unit 4 that the score is

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{X}^T \text{diag}\left(\frac{\partial \mu_i / \partial \eta_i}{\text{Var}[y_i]}\right) (\boldsymbol{y} - \boldsymbol{\mu}). \tag{5}$$

Note that

$$\frac{\partial \mu_i / \partial \eta_i}{\text{Var}[y_i]} = \frac{\partial \mu_i / \partial \theta_i}{\text{Var}[y_i]} = \frac{\ddot{\psi}(\theta_i)}{\text{Var}[y_i]} = m_i. \tag{6}$$

Therefore, the score equations are

$$0 = \boldsymbol{X}^T \text{diag}\left(m_i\right)\left(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\right) \quad \Longleftrightarrow \quad \sum_{i=1}^{n} m_i x_{ij}(y_i - \widehat{\pi}_i) = 0, \quad j = 0, \ldots, p-1. \tag{7}$$

We can solve these equations using IRLS. The Fisher information is

$$\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}, \quad W_{ii} = \frac{(\partial \mu_i / \partial \eta_i)^2}{\text{Var}[y_i]} = \frac{\ddot{\psi}(\theta_i)^2}{\text{Var}[y_i]} = m_i^2 \text{Var}[y_i] = m_i \pi_i (1 - \pi_i). \tag{8}$$

Wald inference. Using the results in the previous paragraph, we can carry out Wald inference based on the normal approximation

$$\widehat{\boldsymbol{\beta}} \dot{\sim} N\left(\boldsymbol{\beta}, \left(\boldsymbol{X}^T \text{diag}(m_i \widehat{\pi}_i (1 - \widehat{\pi}_i)) \boldsymbol{X}\right)^{-1}\right). \tag{9}$$

TBD: Hauck-Donner effect.

Likelihood ratio inference.

Goodness of fit tests.

Fisher's exact test.

Perfect separability.

2 Poisson regression

3 Negative binomial regression