Unit 5: Generalized linear models: Special cases

Eugene Katsevich

November 16, 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,

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

Thus, conditioning on retrospective sampling changes only the intercept term, but preserves the coefficients of $\boldsymbol{x}$. Therefore, we can carry out inference for $\boldsymbol{\beta}_{-0}$ 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

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

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

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

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

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

This approximation holds for $\sum_{i=1}^n m_i \to \infty$. Unfortunately, Wald inference in finite samples does not always perform very well. The Wald test above is known to be conservative due to the *Hauck-Donner effect*. As an example, consider testing $H_0 : \beta_0 = 0.5$ in the intercept-only model

$$ny \sim \mathrm{Bin}(n, \pi); \quad \mathrm{logit}(\pi) = \beta_0. \tag{10}$$

The Wald test statistic is $z \equiv \widehat{\beta}/\mathrm{SE} = \mathrm{logit}(y)\sqrt{ny(1-y)}$. This test statistic actually tends to *decrease* as $y \to 1$, since the standard error grows faster than the estimate itself. For example, take $n = 25$. Then, $z = 3.3$ for $n = 23/25$ but $z = 3.1$ for $n = 24/25$. So the test statistic becomes less significant as we go further away from the null!

Perfect separability. If we have a situation where a hyperplane in covariate space separates observations with $y_i = 0$ from those with $y_i = 1$, we have *perfect separability*. It turns out that some of the maximum likelihood estimates are infinite in this case. The Wald test completely fails in this case, since it uses the parameter estimates as test statistics.

Likelihood ratio inference. Let's first compute the deviance of a logistic regression model. We have

$$L(\boldsymbol{y}; \boldsymbol{\pi}) = \sum_{i=1}^n m_i y_i \log \pi_i + m_i(1 - y_i)\log(1 - \pi_i), \tag{11}$$

so

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\pi}}) = 2(L(\boldsymbol{y}; \boldsymbol{y}) - L(\boldsymbol{y}; \widehat{\boldsymbol{\pi}})) = 2\sum_{i=1}^n \left(m_i y_i \log \frac{y_i}{\widehat{\pi}_i} + m_i(1 - y_i)\log \frac{1 - y_i}{1 - \widehat{\pi}_i}\right). \tag{12}$$

Letting $\widehat{\boldsymbol{\pi}}_0$ and $\widehat{\boldsymbol{\pi}}_1$ be the MLEs from two nested models, we can then express the likelihood ratio statistic as

$$T^{\mathrm{LRT}} = 2(2(L(\boldsymbol{y}; \widehat{\boldsymbol{\pi}}_1) - L(\boldsymbol{y}; \widehat{\boldsymbol{\pi}}_0))) = 2\sum_{i=1}^n \left(m_i y_i \log \frac{\widehat{\pi}_{i1}}{\widehat{\pi}_{i0}} + m_i(1 - y_i)\log \frac{1 - \widehat{\pi}_{i1}}{1 - \widehat{\pi}_{i0}}\right). \tag{13}$$

We can then construct a likelihood ratio test in the usual way. Likelihood ratio inference can give nontrivial conclusions in cases when Wald inference cannot, e.g. in the case of perfect separability. Indeed, suppose that

$$m_i y_i \sim \mathrm{Bin}(m_i, \pi_i), \quad \mathrm{logit}(\pi_i) = \beta_0 + \beta_1 x_i, \quad i = 1, 2. \tag{14}$$

We would like to test $H_0 : \beta_1 = 0$. Suppose that we observe $(x_1, y_1) = (0, 0)$, $(x_2, y_2) = (1, 1)$, giving us complete separability. Can we still get a meaningful test of H_0? We can write out the likelihood ratio test statistic, which is

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\pi}}) = 2 \left(m_1 \log \frac{1}{1 - \frac{m_2}{m_1 + m_2}} + m_2 \log \frac{1}{\frac{m_2}{m_1 + m_2}} \right) = 2 \left(m_1 \log \frac{m_1 + m_2}{m_1} + m_2 \log \frac{m_1 + m_2}{m_2} \right).$$

This is a number that we can compare to the χ_1^2 distribution to get a p-value, as usual.

Goodness of fit tests. We can test goodness of fit in the grouped logistic regression model by comparing the deviance statistic (12) to the asymptotic null distribution χ_{n-p}^2. We can alternatively use the score test, which gives us Pearson's X^2 statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i (1 - \widehat{\pi}_i)/m_i}. \tag{15}$$

Fisher's exact test.

2 Poisson regression

3 Negative binomial regression