

STAT347: Generalized Linear Models

Lecture 5

Today's topics: Chapters 5.1 - 5.2

- Binary data model: data input, link function
- Application scenarios: 2×2 table, case-control study, classification

1 Binary/binomial data model

If the observation y_i is binomial

$$y_i \sim \text{Binomial}(n_i, p_i)$$

and probability function:

$$f(y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} = \binom{n_i}{y_i} \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i)^{n_i}$$

If $n_i = 1$, then y_i is a 0/1 binary data point (follows a Bernoulli distribution).

1.1 Data input

If X_i are categorical variables, then we may have samples with the same X_i and we can group them together.

- ungrouped data: each $n_i = 1$ and some samples have the same X_i , thus they share the same p_i
- a grouped sample \tilde{y}_k for group k contains n_k ungrouped samples whose X_i are the same we only have group level covariates). As $p_i = g^{-1}(X_i^T \beta)$, samples within the same group share the same mean. Let $I_k = \{i : i \text{ in group } k\}$ be the set of individual binary samples and let $n_k = |I_k|$. Then the response for the group samples is:

$$\tilde{y}_k = \sum_{i \in I_k} y_i \sim \text{Binomial}(n_k, p_k)$$

- The grouped data follows the Binomial distribution because we assume that the samples are independent within each group.
- If there are some unmeasured group-level covariates that affect all samples in the group, it can bring in extra dependency and an inflated variance of \tilde{y}_i . (we will discuss this issue later in detail in Chapter 8 and 9.)

- Let $N = \sum_k n_k$. The likelihood for the ungrouped data is:

$$\begin{aligned} f(y_1, y_2, \dots, y_N) &= \prod_i p_i^{y_i} (1 - p_i)^{1 - y_i} \\ &= \prod_k \prod_{i \in I_k} p_k^{\tilde{y}_k} (1 - p_k)^{n_k - \tilde{y}_k} \end{aligned}$$

The likelihood for the corresponding grouped data is:

$$f(\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_K) = \prod_k \binom{n_k}{y_k} \prod_{i \in I_k} p_k^{\tilde{y}_k} (1 - p_k)^{n_k - \tilde{y}_k}$$

The likelihood is not the same between the grouped data and ungrouped data. However, the log-likelihood function only differs by a constant, thus the GLM solution does not change.

1.2 Link function

The expectation of each sample is $\mathbb{E}(y_i) = n_i p_i$ where n_i is a known constant. Thus we define the link function as a function of p_i

$$g(p_i) = X_i^T \beta$$

Equivalently,

$$p_i = g^{-1}(X_i^T \beta) \in [0, 1]$$

If g is a one-to-one mapping (otherwise there can be identifiability issues) and continuous function, then g^{-1} should be monotone. In that case, one natural choice of g^{-1} is to make it as a cdf of some distribution. We then can denote $F(z) = g^{-1}(z)$ as some cdf function. Let $\epsilon_i \stackrel{i.i.d.}{\sim} F(\cdot)$

$$p_i = F(X_i^T \beta) = \mathbb{P}(\epsilon_i \leq X_i^T \beta) = \mathbb{P}(X_i^T \beta - \epsilon_i \geq 0)$$

If y_i is binary, this indicates that y_i follows the distribution

$$Y_i = \begin{cases} 1 & \text{if } X_i^T \beta - \epsilon_i \geq 0 \\ 0 & \text{else} \end{cases}$$

This is also called a latent variable threshold model.

Popular latent variable threshold models:

- The probit link: $F(z)$ is the cdf of a standard Gaussian distribution

$$p_i = \mathbb{P}(X_i^T \beta - \epsilon_i \geq 0) = \mathbb{P}(X_i^T \beta + \epsilon_i \geq 0)$$

where $\epsilon_i \sim N(0, 1)$. Let the hidden variable be $y_i^* = X_i^T \beta + \epsilon_i$, then it goes to the definition of the probit link that some of you may be more familiar with:

$$Y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{else} \end{cases}$$

- The logit link: $F(z)$ is the cdf of a standard logistic distribution

$$F(z) = \frac{e^z}{1 + e^z}$$

- The link function is called the logit link: $g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$
- The logit link is the canonical link of the Binomial distribution
- The identity link: $F(z)$ is the cdf of a uniform $[0, 1]$ distribution and $p_i = X_i^T \beta$
 - The identity link corresponds to a uniform cdf only when $X_i^T \beta \in [0, 1]$ for all samples.
 - Because of the range issue, when using R to solve a binomial GLM with identity link, there can often be numerical problems (such as the error we saw in the earlier data example in Section 1.4, Data Example 1).
- The log-log link: $F(z)$ is the cdf of a standard double-exponential distribution (Gumbel distribution)

$$F(z) = e^{-e^{-z}}$$

- The link function is called the log-log link:

$$g(p_i) = -\log[-\log(p_i)] = X_i^T \beta$$

- Both the probit and logit link assumes a symmetric ϵ_i (around 0). So we implicitly assumed that the response curve is symmetric at 0.5

$$g(p_i) = -g(1 - p_i)$$

One can use the log-log link if such assumption is severely violated (or use a complementary log-log link depending on the shape of the response curve). Read Chapter 5.6.3 for more details (also discussed how one may choose an appropriate link function in practice).

2 Some applications of a Binary GLM

2.1 2×2 table

When Both the X_i and y_i are binary, the grouped data can be represented by a 2×2 table.

- Number of grouped samples: 2.
- Number of total ungrouped observations: $N = n_1 + n_2$ (Table 5.2 of the Agresti book)
- Assume that (X_i, y_i) are i.i.d. Odds ratio (OR) for the response variable Y :

$$\text{OR} = \frac{\mathbb{P}(Y = 1 \mid X = 1)/\mathbb{P}(Y = 0 \mid X = 1)}{\mathbb{P}(Y = 1 \mid X = 0)/\mathbb{P}(Y = 0 \mid X = 0)}$$

- Interpretation of the coefficient β_1 in the binary GLM with logit link: $\text{logit}(p_i) = \beta_0 + \beta_1 X_i$

$$e^{\beta_1} = \text{OR}$$

2.2 Case-control study

We want to know

Risk factor $X \xrightarrow{\text{effect?}}$ Outcome Y

$X_i = 1/0$ if the person is a smoker/non-smoker and $y_i = 1/0$ if the person develops cancer/is a healthy control.

- Prospective design: randomly select smokers and non-smokers from the population and observe whether they will develop cancer in the future.
 - We can compare $\mathbb{E}(Y = 1 | X = 1)$ with $\mathbb{E}(Y = 1 | X = 0)$
 - Drawbacks: the study takes a long time; lung cancer is a rare disease, may observe very few cancer samples.
- Case-control study (retrospective): We randomly select some samples from patients who develop cancer and some samples from healthy controls. Then, we check whether the person has been a smoker or not.
 - We can now only compare $\mathbb{E}(X = 1 | Y = 1)$ with $\mathbb{E}(X = 1 | Y = 0)$
 - The study takes a shorter time, and we can obtain enough cancer cases.

Why is the case-control study popular?

$$\begin{aligned} \text{OR} &= \frac{\mathbb{P}(Y = 1 | X = 1)/\mathbb{P}(Y = 0 | X = 1)}{\mathbb{P}(Y = 1 | X = 0)/\mathbb{P}(Y = 0 | X = 0)} \\ &= \frac{\mathbb{P}(X = 1 | Y = 1)/\mathbb{P}(X = 0 | Y = 1)}{\mathbb{P}(X = 1 | Y = 0)/\mathbb{P}(X = 0 | Y = 0)} \end{aligned}$$

We can also include other covariates \tilde{X} :

$$\begin{aligned} \text{OR} |_{\tilde{X}=x} &= \frac{\mathbb{P}(Y = 1 | X = 1, \tilde{X} = x)/\mathbb{P}(Y = 0 | X = 1, \tilde{X} = x)}{\mathbb{P}(Y = 1 | X = 0, \tilde{X} = x)/\mathbb{P}(Y = 0 | X = 0, \tilde{X} = x)} \\ &= \frac{\mathbb{P}(X = 1 | Y = 1, \tilde{X} = x)/\mathbb{P}(X = 0 | Y = 1, \tilde{X} = x)}{\mathbb{P}(X = 1 | Y = 0, \tilde{X} = x)/\mathbb{P}(X = 0 | Y = 0, \tilde{X} = x)} \end{aligned}$$

Thus, we can study estimate the odds ratio of the risk factor from case-control studies.

Thus, building the logistic regression using case-control study samples is the same as building the model using prospective samples:

$$e^{\beta_1} \equiv \text{OR} |_{\tilde{X}=x}$$

2.3 Classification

Binary GLM models can be used for classification.

Some concepts in evaluating the classification result

- A classification table (Table 5.1 of the Agresti book): y v.s. \hat{y}

- Sensitivity (recall, true positive rate, tpr): $P(\hat{y} = 1 \mid y = 1)$
- Specificity: $P(\hat{y} = 0 \mid y = 0)$
- False positive rate (fpr): $1 - \text{specificity} = P(\hat{y} = 1 \mid y = 0)$
- ROC curve (Figure 5.2 of the Agresti book): fpr v.s. sensitivity

Next time: Chapter 5.3 - 5.5, 5.7, binary GLM: inference, model fitting and examples