STATS 205: Hierarchical Linear Model **Fall 2018**

Lecture 7: Logistic Regression and Generalized Linear Models

Lecturer: Prof. Jingyi Jessica Li *Subscribers: Jinshu Li and Shuai Zhu*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

7.1 Logistic Regression for Binary Data

Suppose we want to predict whether an individual will vote Yes or No. For the i^{th} individual, let:

$$Y_i = \begin{cases} 0, & \text{if he/she voted No} \\ 1, & \text{if he/she voted Yes} \end{cases}$$

Some possible predictors are age, education, and salary. Predictors can be either numerical, or categorical. In this case age and salary would be numerical, while education would be categorical. Given these predictors, we want to predict whether a voter will vote Yes or No.

7.1.1 Random and Systematic Components

We can break up the the logistic regression into random and systematic components. Then, let the random structure be

$$Y_i \sim Bernoulli(\pi_i)$$

where its expectation and variance are

$$\mathbb{E}[Y_i] = \pi_i$$
$$\text{Var}[Y_i] = \pi_i(1-\pi_i)$$

Here, π_i is the probability of success and is between 0 and 1. Note that unlike linear regression, the variance is not constant.

Remember for linear regression the systematic structure follows

$$\pi_i = x_i^T\beta \qquad (Proposal\ One)$$

where x_i^T is a p-dimensional row vector containing the p predictors' values of the i-th observation, and β is a p-dimensional column vector containing the p predictors' coefficients. Note that the first "predictor" is the constant (or intercept).

However, we run into a problem if we want to use this structure. π_i has a range from 0 to 1, while $x_i^T\beta$ has a range of $-\infty$ to ∞. Notice that the range of $x_i^T\beta$ exceeds the range of π_i.

Define the odds to be

$$\frac{\pi_i}{1-\pi_i} = \frac{P(Y_i=1)}{P(Y_i=0)} = x_i^T\beta \qquad (Proposal\ Two)$$

This has a range from 0 to ∞, but still does not match the range of $x_i^T\beta$ ($-\infty$ to ∞).

Therefore, we further modify this to

$$\eta_i = logit(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$$

which has a range from ($-\infty$ to ∞), matching the range of $x_i^T\beta$.

Hence, a reasonable systematic structure is

$$\eta_i = log(\frac{\pi_i}{1-\pi_i}) = x_i^T\beta \qquad (Proposal\ Three)$$

which is often called the linear predictor.

Features of $logit(\pi_i)$:

- $logit(\pi_i)$ is a monotone increasing function.
- If $logit(\pi_i)$ is negative, then the odds are less than $\frac{1}{2}$.
- Conversely, if $logit(\pi_i)$ is positive, then the odds are greater than $\frac{1}{2}$.
- β_j measures the increase in $logit(\pi_i)$ when x_{ij} increases by 1.

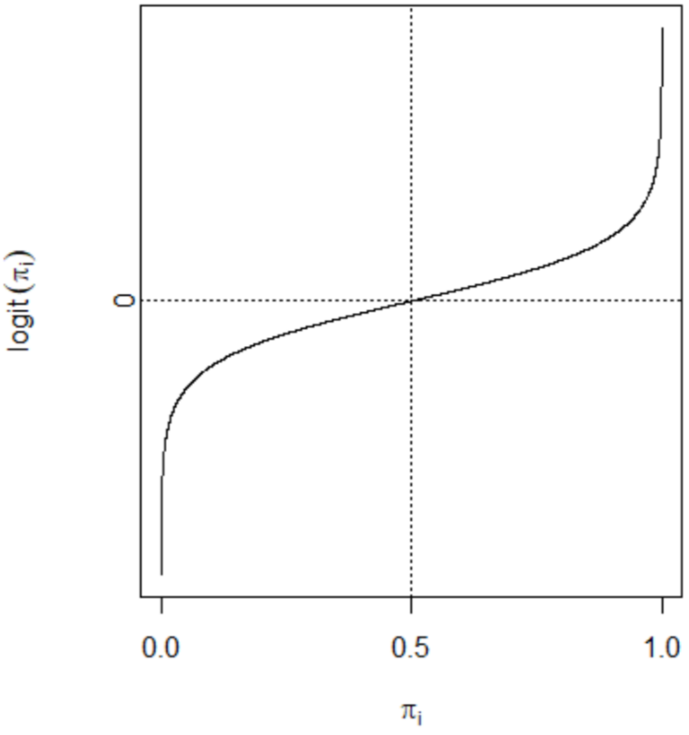

Figure 7.1: Logit function

Remark: In the above logistic regression model, there is no obvious error term. For example, in the linear model $Y_i = \mu_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. However, the logistic model, $Y_i \sim Bernoulli(\pi_i)$, doesn't have an error term. In fact, the randomness of Y_i comes from the Bernoulli distribution.

7.1.2 Grouping Observations by Predictor Values

In practice, sometimes all of the predictors are categorical factors. In these cases, some observations will have the exact same values in all predictors. Then, we group the observations into n groups, where n is the number of distinct observations.

n: Number of groups
n_i: Number of the observations in the i^{th} group

$$Y_{ij} \sim Bernoulli(\pi_i)\,, \text{ where } \; j = 1, \ldots, n_i \text{ and } i = 1, \ldots, n.$$

Each observation in the same group will have the same Bernoulli distribution. Further, we can calculate $Y_i = \sum_{j=1}^{n} Y_{ij} \sim Binomial(n_i, \pi_i)$. Y_i will be the count of the 1's (Yes) in the i^{th} group.

Note: An alternative way to view the relationship between π_i and x_i is

$$\pi_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}.$$

$e^{x_i^T \beta}$ is never negative, so π_i must be between 0 and 1. We are interested in the effect on π_i whenever x_{ij} increases by 1. To find this, we simply take the partial derivative of π_i against x_{ij}:

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \pi_i (1 - \pi_i) \qquad \text{(HW problem, Note } \beta_j \text{ is effect size)}$$

Compared to the linear model, where the change in μ_i is calculated as

$$\frac{\partial \mu_i}{\partial x_{ij}} = \beta_j.$$

Note that in the linear model, the effect on μ_i depends only on the coefficient, but in the logistic regression, the effect on π_i depends on both the coefficient and π_i itself. If π_i is too big/small, it won't cause effect.

7.1.3 Estimation of β

We will use the maximum likelihood estimation approach to estimate β. We will first write down the likelihood function of π.

$$L(\pi) = f(y_i, \ldots, y_n | \pi) \overset{ind}{=} \prod_{i=1}^{n} f(y_i | \pi_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i},$$

which is the joint probability mass function of n Bernoulli variables.
(Note: If we use grouped data, then the equation becomes: $L(\pi) = f(y_i, \ldots, y_n|\pi) \overset{ind}{=} \prod_{i=1}^{n} f(y_i|\pi_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$).

Then the log likelihood is

$$\ell(\pi) = \log L(\pi) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)] = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log(1 - \pi_i) \right]$$

To find β, we use the relationship between π_i and β: $\pi_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$, and then write the log-likelihood in terms of β:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i x_i^T \beta - \log(1 + e^{x_i^T \beta}) \right].$$

Now, we want to find the maximizer β.

$$\hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmax}}\ l(\beta)$$

It is difficult to calculate the derivative $\frac{\partial \ell(\beta)}{\partial \beta_j} = 0$. Actually there is no close form solution for estimating $\hat{\beta}$. Therefore, we will use an optimization algorithm to find the maximum $l(\beta)$ value.

7.2 Generalized Linear Model (GLM) Theory

Linear Model:

Random:

$$Y_i \overset{iid}{\sim} N(\mu_i, \sigma^2)$$

Systematic:

$$\mu_i = x_i^T \beta$$

There are two aspects of the generalization of linear models.

7.2.1 Generalization of the normal distribution

Exponential family density

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

Notes

- $a_i(\cdot)$, $b(\cdot)$, $c(\cdot, \cdot)$ are known functions
- θ_i is the location (shift, related to the mean) parameter
- ϕ is the scale (related to the variance) parameter

Example:

$$Y_i \overset{iid}{\sim} N(\mu_i, \sigma^2)$$

$$\begin{align} f(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\} \tag{7.1} \\ &= \exp\left\{\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \tag{7.2} \end{align}$$

Therefore, $\theta_i = \mu_i, \quad \phi = \sigma^2, \quad b(\theta_i) = \frac{1}{2}\theta_i^2, \quad a_i(\phi) = \phi$

Theorem 7.1 (Properties of the exponential family density) *If $Y_i \sim f(y_i; \theta_i, \phi)$, then*

- $E(Y_i) = b'(\theta_i)$;
- $Var(Y_i) = b''(\theta_i) \cdot a_i(\phi)$;

Proof:

$$\int f(y_i)\mathrm{d}y_i = 1 \implies \tfrac{\partial}{\partial\theta_i}\int f(y_i)\mathrm{d}y_i = 0 \overset{exchange}{\implies} \int \frac{\partial}{\partial\theta_i} f(y_i)\mathrm{d}y_i = 0$$

(the exchange of differentiation and integration is valid for f under the exponential family, smoothness in f)

$$\implies \int \frac{y_i - b'(\theta_i)}{a_i(\phi)} f(y_i)\mathrm{d}y_i = 0 \implies \int y_i f(y_i)\mathrm{d}y_i = b'(\theta_i)\int f(y_i)\mathrm{d}y_i$$

$$\implies E(Y_i) = b'(\theta_i)$$

Also, $\tfrac{\partial^2}{\partial\theta^2}\int f(y_i)\mathrm{d}y_i = 0, \overset{exchange}{\implies} \int \frac{\partial^2}{\partial\theta^2} f(y_i)\mathrm{d}y_i = 0$

$$\implies [\int \frac{-b''(\theta_i)}{a_i(\phi)} f(y_i) + \left(\frac{y_i - b'(\theta_i)}{a_i(\phi)}\right)^2 f(y_i)]\mathrm{d}y_i = 0$$

$$\implies \frac{1}{a_i(\phi)^2} Var(Y_i) = \frac{b''(\theta_i)}{a_i(\phi)} \qquad \text{(since } E(Y_i) = b'(\theta_i))$$

$$\implies Var(Y_i) = b''(\theta_i) a_i(\phi)$$

7.2.2 Generalization of the systematic structure

Link function: monotone increasing and differentiable function g, which links $\mu_i = E(Y_i)$ with the linear predictor

$\eta_i = g(\mu_i) = x_i^T\beta \iff \mu_i = g^{-1}(\eta_i) = g^{-1}(x_i^T\beta)$

Example:

Logistic regression: $g(\mu_i) = \log(\dfrac{\mu_i}{1-\mu_i}) \iff \mu_i = g^{-1}(x_i^T\beta) = \dfrac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$

Remark:
The two models below are not the same model:

$$\begin{cases} g(Y_i) = x_i^T\beta + \epsilon_i \\ g(\mu_i) = x_i^T\beta \end{cases} \implies \begin{cases} E[g(Y_i)] = x_i^T\beta \\ g[E(Y_i)] = x_i^T\beta \end{cases} \quad \text{while } E[g(Y_i)] \neq g[E(Y_i)]$$

Unless g is linear.