Module 11: Logistic Regression

Jeffrey Woo

MSDS, University of Virginia

Welcome

- Remind me to record the live session!
- Recommended: put yourself on mute unless you want to speak.
- Reminder: the raise hand button can be found under “Manage Participants”.

Agenda

- Q&A (from Module 10, 11)
- A few comments about Module 11
- Small group discussion of guided question set
- Large group discussion of guided question set and other questions that popped up

Comment about Indicator Variables in R

- When using lm() or glm(), R converts factors to indicator variables. So, if your categorical variables are already coded as 0/1 indicator variables, they can be left alone.
- However, for other functions, 0/1 indicator variables may not work. Have to check documentation on how functions handle categorical variables.
- For example, using `glmnet()` for shrinkage methods and `tree()` for tree based methods.
- Most visualizations need categorical variables as factors and cannot be 0/1 coded.

If something in your output doesn't look right and you have categorical variables, make sure that your categorical variables are the correct type.

Logistic Regression Equation

Typically written as

$$\begin{aligned} \log \frac{\pi}{1-\pi} &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \\ &= \boldsymbol{X\beta} \end{aligned}$$

Alternate way of writing:

$$\begin{aligned} \pi &= \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \\ &= \frac{\exp(\boldsymbol{X\beta})}{1 + \exp(\boldsymbol{X\beta})} \end{aligned}$$

Probability against X

Parameter Estimation

- For linear regression, the parameters can be estimated by least squares criterion, and we have a closed form solution for the vector of estimated β's:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}$$

- For logistic regression, we cannot use least squares, since least squares requires a response variable that is quantitative.
- Instead, the method of maximum likelihood is used.

Method of Maximum Likelihood

General idea: model parameters are found by maximizing the likelihood that the process described by the model produced the observed data.

Method of Maximum Likelihood

Since we assumed the response variable follows a Bernoulli distribution, the probability distribution of each observation is

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{1}$$

Note: y_i is either 0 or 1.

- Probability distribution: mathematical function that gives the probabilities of occurrence of different possible outcomes for a random variable.

Method of Maximum Likelihood

Since we assumed the observations are independent, the likelihood function is

$$L(y_1, \cdots, y_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{2}$$

We want to maximize the likelihood function (2) with respect to $\boldsymbol{\beta}$. Since maximizing a function is the same as maximizing the log of a function, we actually maximize the log-likelihood function

$$\log L(y_1, \cdots, y_n, \boldsymbol{\beta}). \tag{3}$$

Method of Maximum Likelihood

Using $\pi_i = \frac{\exp(\boldsymbol{x_i'}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x_i'}\boldsymbol{\beta})}$ and after some algebra on the log-likelihood function (3), we maximize the following with respect to $\boldsymbol{\beta}$

$$\log L(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \boldsymbol{x_i'}\boldsymbol{\beta} - \sum_{i=1}^{n} \log[1 + \exp(\boldsymbol{x_i'}\boldsymbol{\beta})] \tag{4}$$

Method of Maximum Likelihood

There is no closed-form solution to maximizing (4). Numerical method called iteratively reweighted least squares (Newton's method, Fisher scoring) is used.

- Method of maximum likelihood can also be used in linear regression. It turns out that estimators from both approaches for linear regression converge to one another.

Reproducibility Poll

Run these 2 lines of code and record the 5 integers you get

```
set.seed(1)
sample.int(100,5)
```

Small Group Discussion

- Materials can be found under Module 11 Live session.
- Have the guided question set and corresponding data set open.
- Have R open.
- Recommended: have easy access to your notes, textbook, as well as the tutorial.
- You can see who your group members are. As well as some roles you will have in your small group. Roles will rotate each session.

Goodness of Fit Tests

Three Goodness of Fit (GOF) tests in logistic regression. In other words is the log odds a linear combination of coefficients and predictors:

1. Deviance GOF test
2. Pearson GOF test
3. Hosmer-Lemeshow test

Deviance and Pearson require grouped data (usually found in designed experiments); cannot have (a number of) observations that have unique combination of predictors. Difficult to implement in observational studies.

Residuals in Logistic Regression

Due to the binary nature of the response variable in logistic regression, examining residuals is not very helpful (due to discrete nature of response, the values of the residuals are typically discrete). Again, require data are grouped to reliably use the plots.

Where Are We Headed?

- Module 12: Validating the Logistic Regression Model.

Upcoming

- Next Tuesday, Nov 29: Module 12.
- I've combined HW 11 & 12 into one HW, due Dec 5. You can work on a majority of these questions already.
- Exam, due Dec 4. Covers Modules 3 to 9. Opens Nov 29 after our live session.
- Project 2 due Dec 14 (bulk of work) and Dec 15.