Logistic Regression

insMany of the formulas are primarily derived from *Linear Regression Analysis, Fifth Edition* by Montgomery et al. and/or the materials from the STAT 6021 course taught by Dr. Woo at the University of Virginia. It is believed that the materials are well-known equations and concepts in the public domain. If you believe otherwise, please reach out to me through my Github account so that I can correct the material. If not otherwise stated, quotes are from the textbook.

Warning!

This is a work in progress. It is still being corrected and revised. Please consult the textbook to verify formulas and propositions in this text. If you find an error, let me know.

Introduction

Logistic regression is a way to model binary responses (0 or 1) using regression analysis under appropriate conditions. It is covered in chapter 13 of the textbook.

In normal regression analysis, we model the response in terms of a linear combination of predictors:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

This model does not work well for cases when the response varies between two values with a fixed probability (i.e., is a Bernoulli random variable). Instead, we take a different approach.

Definitions

Trials and observations

- The response is a binary outcome, 1 or 0. It is not atypical to pick one of these levels (usually 1) to represent a successful outcome, a postive outcome, or some other kind of outcome.
- A trial is a particular measurement of the response for a given set of levels (i.e., values) for the predictors.
- An observation is a measurement of the number of 1's obtained by a series of trials at the same levels for the predictors. The book uses n_i and y_i to represent the number of trials and number of 1's in those trials respectively. For example:

observation (i)	x levels (X_i)	response (y)	trials	no. trials (n_i)	result (y_i)
1	Carl flips coins	landing face (H = 1)	H H T	3	2
2	Sam flips coins	landing face (H = 1)	T H	2	1

Odds and sigmoid functions

- The *odds of an event* is the probability of it happening divided by the probability of it not happening. For example, if you have a 30% chance of winning at Blackjack, then your odds are

$$\frac{30}{100} = \frac{3}{10} = 0.3$$

- The *log-odds* of an event are the (natural) logarithm of the odds of the event:

$$\text{log-odds} = \ln(\text{odds})$$

- A Bernoulli random variable represents an event that has two possible outcomes (occurring or not), with a fixed probability π. The typical example is flipping a coin to see if you get heads. Another example is the event of drawing a king of hearts after reshuffling a deck.
- The *logistic* and *logit* transforms respectively are

$$\text{logistic}(x) = \frac{x}{1+x} \qquad \text{logit}(x) = \frac{x}{1-x},$$

where the two are inverses:

$$\text{logistic}(\text{logit}(x)) = x.$$

- The *logstistic* transform is part of a broader class of common transforms called *sigmoid* functions. They have this name because they make curves that look like the letter *s*. Sigmoid literally means "shaped like the letter S" (or if you know your Greek, ς).

Basic Model

Suppose that each response y_i is a Bernoulli random variable with probability π_i:

$$y = \begin{cases} 1 & \text{with probability } \pi_i \\ 0 & \text{with probability } 1-\pi_i \end{cases}$$

Theoretically, the log-odds η of y_i are

$$\eta_i = \ln \frac{\pi_i}{1-\pi_i},$$

where $f(x) = x/(1-x)$ is known in general as the *logit* transform.

To estimate the log-odds in terms of predictors, we estimate η using a linear model :

$$\hat{\eta}_x = \mathbf{x}'\boldsymbol{\beta},$$

which we can call the *linear predictor* (for the log-odds) given the specific predictor level. Since we are interested in the odds calculated for specific levels of $\mathbf{x}$ here, I have added a subscript. Outside of this section, I will sometimes drop the subscript for notational convenience. The book uses i as a subscript to indicate the result for the $\mathbf{x}$ level derived from observation i.

Since the logistic transform inverts the logit transform, we can recover π_x using the logistic transform on the log-odds:

$$\hat{\pi}_x = \frac{1}{1 + e^{\hat{\eta}_x}}.$$

We offer without proof that $\hat{\pi}_X$ estimates $E(\text{response}|\mathbf{x})$. Supposing that result, we have successfully constructed a model for our response! We can therefore express the expected response in terms of the predictor:

$$\hat{y}(\mathbf{x}) = \hat{\pi}_x = \frac{1}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}},$$

where $\hat{y}$ here is a fitted value that we use as our estimate of the average response for the given level of $\mathbf{x}$.

Estimates

Calculating

We use the maximum-likelihood estimates to calculate values for the $\hat{\boldsymbol{\beta}}$. These are typically obtained using numerical methods (e.g., with a computer). The book suggests using iteratively reweighted least squares.

For interest, the likelhood function is

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \pi^{y_i}(1 - \pi_i)^{n_i - y_i}.$$

In some places, the book assumes that $n_i = 1$ and in others it does not.

We have borrowed some notation from the Wikipedia article *Likelihood function* ((link)[https://en.wikipedia.org/w/index.php?title=Likelihood_function&oldid=923888811]) instead of using the book's somewhat more confusing notation.

The log-likelihood is

$$\ln L(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n} y_i \ln(\pi_i) + \sum_{i=1}^{n} (n_i - y_i) \ln(1 - \pi_i).$$

Properties

The variance matrix is

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1},$$

where the diagonals of $\mathbf{V}$ are

$$V_{ii} = n_i \hat{\pi}_i (1 - \hat{\pi}_i).$$

The expected value is

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},$$

showing that it is not a biased estimator (under our assumptions).

Statistic inference

To quote the textbook:

> "Statistical inference in logistic regression is based on certain properties of maximum-likelihood estimators and on likelihood ratio tests. These are large-sample or **asymptotic** results."

Checking for significant coefficients

Likelihood ratio test

General method

Consider a full model (FM) and a reduced model (RM). Let $L(\text{model})$ mean the value of the likelihood function for a given model. Then let

$$LR = 2ln\frac{L(FM)}{L(RM)}.$$

When the reduced model is correct, this follows a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between models. Therefore, we would reject the null hypothesis that *the reduced model is the better model* if χ^2 test exceeds our level of significance.

Logistic regression

In logistic regression, we use the model under consideration as the full model and the model with only a constant term (ignoring all predictors) as the reduced model. To show that our model is worth considering, we want to reject the constant term model.

Our null hypothesis in this case is $H_0 : \beta_i = 0$ for all $i > 0$.

The formula is

$$LR = 2\left\{\frac{\text{full model}}{(\sum_{i=1}^{n} y_i \ln\hat{\pi}_i) + (\sum_{i=1}^{n}(n_i - y_i)\ln(1-\hat{\pi}_i))} - \frac{\text{reduced}}{[y\ln(y) + (n-y)\ln(n-y) - n\ln(n)]}\right\}.$$

Checking for a good fit

Deviance

The deviance can be used to assess goodness of fit.

The textbook defines the deviance as

$$D = 2\ln\frac{L(\text{saturated model})}{L(\text{full model})}.$$

In this case, the deviance is

$$D = 2\sum_{i=1}^{n}\left[y_i \ln\frac{y_i}{n_i\pi_i} + (n_i - y_i)\ln\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}\right].$$

For large enough samples and an adequate fit by the full model, the deviance follows a χ^2 distribution with $n-p$ degrees of freedom, where p is the number of coefficients in the model ($p = k+1$ for k many predictors).

Note by JG

The book indicates that the saturated model uses $\pi_i = y_i/n_i$. Comparing this to the above formula, it works out. The saturated model is:

$$\ln L(\text{saturated}) = \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{n_i}\right) + \sum_{i=1}^{n}(n_i - y_i)\ln\left(1 - \frac{y_i}{n_i}\right).$$

Pearson chi-square

There is also a Pearson chi-square goodness-of-fit statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{y_i - n_i\hat{\pi}_i}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

This follows a χ^2 distribution with $n-p$ degrees of freedom. In addition to seeing if the value is large (according to its distribution), you can divide by $n-p$ and compare the value to unity to assess the goodness of fit.

Homer-Lemeshow test

This can be used if there are no repeated instances of $\mathbf{x}$ levels.

First, group the observations into g groups (normally 10). Then define

- O_j = the number of successes in group j
- N_j = the number of observations in group j
- $\bar{\pi}_j$ the average number probability of success in group j:

$$\bar{\pi}_j = \sum_{i \in \text{group j}} \frac{\hat{\pi}_j}{N_j}.$$

Then the statistic is

$$HL = \sum_{j=1}^{g} \frac{(O_j - N_j\bar{\pi}_j)^2}{N_j\bar{\pi}_j(1-\bar{\pi}_j)}.$$

Under the assumption that the fitted model is correct, the statistic follows a χ^2 distribution with $(g-2)$ d.f.

Note by JG

I think the book has a misprint in the formula for HL. It uses $j \in [1, n]$. I have modified this.

Testing subsets of parameters

First split the estimated parameters:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2.$$

Then we have the hypothesis test

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}. \qquad H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}.$$

In this case, the reduced model is

$$\eta_{\text{reduced}} = \mathbf{X}_1\boldsymbol{\beta}_1.$$

We consider the deviance (also called the partial deviance)

$$D(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = D(\boldsymbol{\beta}_1) - D(\boldsymbol{\beta}).$$

If the null hypothesis holds and for large n, this follow a chi^2 distribution with r degrees of freedom, where r is the number of coefficients we removed to make the reduced model.

If $D(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) \geq \chi^2_{a,r}$, we reject the null hypothesis. Otherwise, we do not reject the null hypothesis.

Tests on individual coefficients

You can use the subset of parameters approach above. There's another option. We can use Wald inference. This requires a large enough sample size.

Consider the hypothesis:

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0$$

Then

$$Z_0 = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$$

follows the standard normal distribution.

You can calculate the standard errors by getting the variance matrix, picking out the appropriate value, and then taking the square root.

Textbook pages for other topics

We present some page numbers for sections not covered here.

Model adequacy

See p. 440.

Other models

See p. 442.

More than two possible outcomes

See p. 442.

See the textbook for details.