(Binary) Logistic Regression Reference Sheet

Vincent La

May 29, 2017

1 Key Terms

- **dichotomous variable** - A variable which takes on one of two values, usually 0 (no/negative) or 1 (yes/positive). *Synonyms: Binary variable*
- **odds** - The probability that something will happen divided by the probability it will not. For example, if an event has a $\frac{2}{3}$ chance of happening and a $\frac{1}{3}$ chance of not, then its odds of occurring are $\frac{2}{3}/\frac{1}{3} = 2$ or $2:1$.
- **covariates** - The predictor variables in a regression model. Also referred to as independent variables, regressors, explanatory variables, risk factors (in medical literature), etc...

2 Goal

Suppose we have some binary outcome Y, e.g. pass or fail, infected or not infected, and some predictors $X_1, ..., X_n$. The goal of logistic regression is to model the relationship between $X_1, ..., X_k$ and Y.

3 Form of the Logistic Regression Model

More formally, this can be described as follows:

- Data Y_i are Binomial random variables with probability π_i
- π_i is the conditional probability that $Y_i = 1$ given the predictors $X_1, ..., X_n$, i.e. $\pi = \mathbb{P}(Y = 1 | X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$
- The relationship between an individual observation Y_i and its associated predictors is given by:

$$\pi_i(x_1, x_2, ..., x_k) = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_k)}$$

- The **logit transformation** is used to describe the previous relationship as a linear combination of the predictors

$$logit(\pi) = log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + ... + \beta_2 x_2 + ... + \beta_k x_k$$

4 Logistic Regression as a Generalized Linear Model

A generalized linear model is any model which attempts to predict some response as the function of a linear combination of predictors. After applying the logit transformation, logistic regression can be described as a generalized linear model.

link function A function which transforms a response such that it can be written as a linear combination of predictors

4.1 Example: Linear Regression

The linear regression model is written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

In contrast with logistic regression, in linear regression:

- Data: Y_i are continuous real-valued random variables
- Goal: Predict the value of Y as a linear function of the predictors, i.e.

$$\mathbb{E}(Y_i | X_1 = x_1, x_2, ..., x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

The link function in the case of linear regression is the **identity function**, which is simply $y = y$ (i.e. the identity function doesn't really do anything).

4.2 Logistic Regression

The logistic regression model is written as

$$logit(\pi) = log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + ... + \beta_2 x_2 + ... + \beta_k x_k$$

The link function for logistic regression is the logit function, which is the inverse cumulative distribution function of the logistic distribution.

5 Model Fitting

5.1 Maximum Likelihood Estimation

The coefficients of a logistic regression model are found by maximizing the the **likelihood function**

$$\Pi_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{n_i - Y_i}$$

with respect to $\beta_0, \beta_1, ...\beta_k$.

5.2 Contrast: Linear Regression

In the linear regression model, the best fitting β coefficients are the **ordinary least squares (OLS) estimators** given by minimizing the sum of squared residuals with respect to β.

$$\begin{aligned} SSR &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - ... - \hat{\beta}_k X_{ik})^2 \end{aligned}$$

For simple linear regression, we can describe the optimal fitting β coefficients as

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_1 - \bar{x})^2} \end{aligned}$$

and for multiple regression as

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

However, in logistic regression there is no closed form expression for the best fitting β coefficients.

6 Hypothesis Testing and Confidence Intervals

7 Interpretation of Coefficients