Lectures on GLM
Stat 431, Summer 2012

Hyunseung Kang

Aug 1-2 2012

1 Introduction

Consider the spam detection problem we considered at the beginning of the class. Our goal is to determine whether a particular e-mail with a set of features, denoted as $X_{i,1}, ..., X_{i,p}$ where each $X_{i,j}$ represented some feature of the ith e-mail, is spam or not. Let spam mail be denoted as Y_i where $Y_i = 1$ is spam and $Y_i = 0$ is not spam. Based on $(Y_i, X_{i,1}, ..., X_{i,p})$, we want to build a model that predicts Y_i given a set of $X_{i,1}, ..., X_{i,p}$

So far in linear regression, we dealt with the case when Y_i was continuous. Specifically, we assume that Y_i was distributed as a Normal distribution with mean

$$E(Y|X_{,1}, ..., X_{,p}) = \beta_0 + \beta_1 X_{,1} + ... + \beta_p X_{,p} \tag{1}$$

and variance σ^2. However, the spam detection problem provides a unique modeling challenge. Because Y_i is a binary variable, it is unreasonable to assume that Y_i is a continuous distribution, or even Normal. Another way to say it is that it is unreasonable to assume that $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Instead, we need a model of Y_i in the following form

$$P(Y_i = 1|X_{i,1}, ..., X_{i,p}) \tag{2}$$

In this lecture we'll discuss how to model this probability expression using a logistic function.

2 Logistic and Logit Regression

2.1 Model

The model for logistic regression is as follows[1]

$$\pi(\mathbf{X}) \overset{\text{def}}{=} P(Y_i = 1|X_{i,1}, ..., X_{i,p}) = \frac{e^{\beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p}}}{1 + e^{\beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p}}} \tag{3}$$

Equation (3) is bounded between 0 and 1, for any values of $X_{i,j}$ and β_j. You can think of $\pi(\mathbf{X})$ as the expression for the Often in statistical inference and in epidemiology, we look at the *odds ratio*, defined as

$$\frac{P(Y_i = 1|X_{i,1}, ..., X_{i,p})}{P(Y_i = 0|X_{i,1}, ..., X_{i,p})} = \frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})}$$

If the odds are greater than 1, the probability of $Y_i = 1$ is greater than the probability of $Y_i = 0$ and $Y_i = 1$ is more favorable. If the odds are less than 1, the probability of $Y_i = 1$ is less than the probability of $Y_i = 0$ and $Y_i = 0$ is more favorable.

Taking the logarithm of the odds ratio, we end up with our familiar expression

$$\log\left(\frac{\pi(\mathbf{X})}{1 - \pi(\mathbf{X})}\right) = \beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p} \tag{4}$$

[1] We simplify notation by using $\pi(\mathbf{X})$ to represent the probability of $Y_i = 1$ given $X_{i,1}, ..., X_{i,p}$. Here, $\mathbf{X}$ is a vector where $\mathbf{X} = (X_{i,1}, ..., X_{i,p})$

This type of transformation on equation (3) is known as the *logit* transformation. Specifically, logit transformations are transformations where we take a value $0 < z < 1$ and transform it to $\log\left(\frac{z}{1-z}\right)$. The $\log\left(\frac{z}{1-z}\right)$ as a function of z is known as the *link function* .

To estimate β_j from the data, we cannot rely on the sum of squares of residuals. Instead, we must use a likelihood-based approach where we find the β_j that maximize the probability of observing $(Y_i, X_{i,1}, ..., X_{i,p})$ is maximized under the model specified in equation (3). Mathematically, we maximize[2]

$$\max_{\beta_j} \prod_{i=1}^{n} (\pi(\mathbf{X}))^{I(Y_i=1)} (1 - \pi(\mathbf{X}))^{I(Y_i=0)}$$

This maximization of the *likelihood* is easier when we take the log, called the *log likelihood.* Thankfully, R can handle this math and you don't have to worry about it

2.2 Inference

Much like multiple regression, we have inference procedures for a single β_j or a group of β_j's in logistic regression. There are two procedures to do this type of inference. Here, we'll only focus on *likelihood ratio tests*[3]

First, we define deviance to the be the difference in log likelihood of the reduced model and the full model

$$\Delta \overset{\text{def}}{=} -2\log(lk_{reduced}/lk_{full}) \tag{5}$$

where lk is the likelihood of the reduced and the full model. If there is no difference between the full and the reduced model, the deviance is zero. If there is a significant difference between the reduced and the full model, the deviance is greater than zero. For large sample size, equation (5) is distributed as a chi-square distribution with degrees of freedom that is equal to the difference in the number of parameters beween the reduced and the full model.

3 Poisson Regression

Poisson regression deals with the case when Y is a count variable. We have the model

$$E(Y|X_{,1}, ..., X_{,p}) = e^{\beta_0 + \beta_1 X_{,1} + ... + \beta_p X_{,p}}$$

[2] $I()$ is an indicator function that takes on value 1 if the event inside the parenthesis occurred and 0 otherwise

[3] Wald test or Wald t-test is another procedure used frequently in the literature. For large sample size, the Wald test and the likelihood ratio test are indistinguishable. However, for smaller sample size, the likelihood ratio test test tends to perform better.