Journal of the Korean Data & Information Science Society
2011, **22**(1), 107–113

한국데이터정보과학회지

A study on log-density ratio in logistic regression model for binary data[†]

Myung-Wook Kahng[1]

[1]Department of Statistics, Sookmyung Women's University

Received 29 October 2010, revised 29 December 2010, accepted 03 January 2011

Abstract

We present methods for studying the log-density ratio, which allow us to select which predictors are needed, and how they should be included in the logistic regression model. Under multivariate normal distributional assumptions, we investigate the form of the log-density ratio as a function of many predictors. The linear, quadratic and crossproduct terms are required in general. If two covariance matrices are equal, then the crossproduct and quadratic terms are not needed. If the variables are uncorrelated, we do not need the crossproduct terms, but we still need the linear and quadratic terms.

Keywords: Binary regression, kernel mean function, log-density ratio, log-odds ratio, logistic regression.

1. Introduction

Logistic regression models are useful in problems where the dependent variable takes on only a few discrete values. Major fields of application inc1ude econometrics, biostatistics, and educational testing. We consider the special case in which the response is binary. The classical theoretical treatment of binary data is that of Cox (1970). There have been many theoretical developments and extensions associated with its use. See Anderson (1984) and McCullagh (1980) for models for nominal and ordinal categorical response data.

The general goal of a regression analysis is to understand how the conditional cdf $F(y|\boldsymbol{x})$ varies as a set of p predictors $\boldsymbol{x}$ varies. In the case of a binary regression, this cdf is completely characterized by the mean function $E(y|\boldsymbol{x})$.

In this article we introduce some ideas on how to extract relevant statistical information in logistic regression models for binary responses, and we present some applications based on real data sets. After reviewing the general logistic regression context, we introduce the log-density ratio. This forms the building block of our approach, which essentially amounts to studying the inverse problem, that is the conditional distribution of $\boldsymbol{x}$ given y, to guide the model specification for the regression of y on $\boldsymbol{x}$. Under some multivariate distributional assumptions, we investigate the form of the log-density ratio as a function of the predictors.

[†] This research was supported by the Sookmyung Women's University Research Grants 2009.
[1] Professor, Department of Statistics, Sookmyung Women's University, Seoul, 140-742, Korea.
E-mail: mwkahng@sookmyung.ac.kr

2. Logistic model and the log-density ratio

Suppose we observe a binary response variable y, and we want to study its relationship with a set of p predictors $\boldsymbol{x}$. The outcome variable is assumed to be distributed as a Bernoulli random variable with probability of success given by $\theta(\boldsymbol{x})$. For such distribution $E(y|\boldsymbol{x}) = P(y = 1|\boldsymbol{x}) = \theta(\boldsymbol{x})$, so the mean function completely characterizes the stochastic component of the model. The dependence of the mean function on the predictors is expressed through a function of the linear combination of the predictors. For logistic regression models, this can be written as

$$E(y|\boldsymbol{x}) = \theta(\boldsymbol{x}) = \frac{exp(\boldsymbol{\beta^T x})}{1 + exp(\boldsymbol{\beta^T x})}.$$

This is called the kernel mean function (Cook and Weisberg, 1999), and its inverse defines the link function (Nelder and Wedderburn, 1972)

$$log\ \left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \boldsymbol{\beta^T x}.$$

The linear predictor $\boldsymbol{\beta}^T\boldsymbol{x}$ defines the systematic component of the model. In general, this could consist of several terms obtained as functions of $\boldsymbol{x}$, such as transformations, powers, cross-products, indicators for factors, etc. To emphasize this fact, Cook and Weisberg (1999) denote the terms in the systematic component of a model with $u = u(\boldsymbol{x})$, a vector of p' terms derived from the p predictors $\boldsymbol{x}$.

The logistic regression model belongs to the broad class of generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). Maximum likelihood estimates of the parameters in the $\boldsymbol{\beta}$ are available, although not in closed form, so they are usually obtained by an iterative procedure. Since the logit link function is the canonical link for binomial data, the two methods provide the same estimates. It must be noted that there exist other regression models for studying the dependence of a binary variable y on $\boldsymbol{x}$, such as probit regression, and complementary log-log regression function.

3. The log-density ratio

Suppose we have data concerning a binary response variable y taking values 0, 1 and a set of p predictors $\boldsymbol{x} = (x_1,...,x_p)^T$. Denoting the conditional probability density function of $\boldsymbol{x}$ given $y = j$, $j = 0, 1$ by $f(\boldsymbol{x}|y = j)$, consider the assumption that the log ratio of these density functions is a linear function of $\boldsymbol{x}$; that is $log[f(\boldsymbol{x}|y = 1)/f(\boldsymbol{x}|y = 0)] = \alpha_0 + \boldsymbol{\alpha_1^T x}$. Since

$$log\ \left(\frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)}\right) = log\ \left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) - log\ \left(\frac{P(y=1)}{P(y=0)}\right),$$

where $p(\boldsymbol{x}) = \theta(y = 1|\boldsymbol{x})$, we have

$$\begin{aligned} log\ \left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) &= log\ \left(\frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)}\right) + log\ \left(\frac{P(y=1)}{P(y=0)}\right) \\ &= log\ \left(\frac{P(y=1)}{P(y=0)}\right) + \alpha_0 + \boldsymbol{\alpha_1^T x} \\ &= \beta_0 + \boldsymbol{\alpha_1^T x}, \end{aligned}$$

where $\beta_0 = \alpha_0 + log\,[P(y=1)/P(y=0)]$. Thus, if the log ratio of the conditional densities of $\boldsymbol{x}$ given y is a linear function of $\boldsymbol{x}$, the logistic model is the correct model for the conditional distribution of y given $\boldsymbol{x}$.

Extending this idea, suppose now that the log ratio of conditional density functions is a linear function, not of $\boldsymbol{x}$, but of some vector function of $\boldsymbol{x}$, for example $\boldsymbol{g}(\boldsymbol{x})$. Thus $log[f(\boldsymbol{x}|y=1)/f(\boldsymbol{x}|y=0)] = \alpha_0 + \boldsymbol{\beta_1^T g}(\boldsymbol{x})$. Using the same arguments as before we find that

$$log\ \left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \beta_0 + \boldsymbol{\beta_1^T g}(\boldsymbol{x}).$$

Thus the logistic model is appropriate if we use $\boldsymbol{g}(\boldsymbol{x})$ as the vector of explanatory variables rather than $\boldsymbol{x}$. This model is different from the generalized additive model (GEM), $log\ \ (\theta(\boldsymbol{x})/1-\theta(\boldsymbol{x})) = \beta_0 + \sum_j f_j(x_j)$, in which the sum of functions of each explanatory variable replaces $\boldsymbol{\beta_1^T g}(\boldsymbol{x})$.

Let $f(\boldsymbol{x}|y=j)$ be the probability density function for $\boldsymbol{x}$ given $y=j$, $j=0,1$, and let $f(\boldsymbol{x})$ be the marginal density function. Using Bayes' theorem, we can write

$$\theta(\boldsymbol{x}) = P(y=1|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|y=1)P(y=1)}{f(\boldsymbol{x})}$$

where we have renamed the mean function as $\theta(\boldsymbol{x})$, the probability that $y=1$ given the value $\boldsymbol{x}$. We can also write

$$1-\theta(\boldsymbol{x}) = P(y=0|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|y=0)P(y=0)}{f(\boldsymbol{x})}.$$

If we take the logarithm of the ratio of these two quantities, we get the log-odds ratio,

$$\begin{aligned} log\ \left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) &= log\ \left(\frac{P(y=1)}{P(y=0)}\right) + log\ \left(\frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)}\right) \\ &= log\ \left(\frac{P(y=1)}{P(y=0)}\right) + h(\boldsymbol{x}). \end{aligned}$$

The log-odds ratio is therefore equal to the sum of two terms. The first term does not involve $\boldsymbol{x}$, and is marginal log-odds of success. The value of this term depends on the sampling design; for example in a retrospective study this quantity is fixed by design. The second term $h(\boldsymbol{x})$ is the log-density ratio.

3.1. The log-density ratio with one predictor

Suppose $f(\boldsymbol{x}|y)$ follows a known parametric form, then the log-density ratio $h(\boldsymbol{x})$ can be used to derive the terms that are needed in a logistic regression model, and the conditions under which these terms are required. For example, assuming a normal distribution, that is $x|(y=j) \sim N(\mu_j, \sigma_j^2)$ for $j = 0, 1$, the log-density ratio can be written as (see, Cook and Weisberg, 1999; Scrucca and Weisberg, 2004)

$$h(\boldsymbol{x}) = log \left(\frac{\frac{1}{\sqrt{2\pi}\sigma_1} exp \ \left[-(x-\mu_1)^2/2\sigma_1^2\right]}{\frac{1}{\sqrt{2\pi}\sigma_0} exp \ \left[-(x-\mu_0)^2/2\sigma_0^2\right]} \right)$$

$$= \left[log \ \left(\frac{\sigma_0}{\sigma_1}\right) + \frac{1}{2}\left(\frac{\mu_0^2}{\sigma_0^2} - \frac{\mu_1^2}{\sigma_1^2}\right) \right] + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right) x + \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right) x^2.$$

The conditions for the inclusion of the linear and the quadratic term can be read off from the above equation. If $f(x|y)$ is normal with different means and variances then we need to include both a linear and a quadratic term for x in the logistic regression model. The quadratic term is not required if the two conditional distributions have the same variance, whereas the linear component is generally needed, except in the case in which the ratio of the mean over the variance is the same in both groups.

3.2. The log-density ratio with many predictors

With many predictors, the idea of studying the inverse problem is still valid, so in principle we could derive the terms required by a logistic regression model using the same approach we adopted in the one predictor case. Unfortunately, the relationships among the predictors make things potentially more complicated.

When the predictors are conditionally independent given y the log-odds can be written as

$$log \ \left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = log \ \left(\frac{P(y=1)}{P(y=0)}\right) + \sum_{i=1}^{p} log \ \left(\frac{f(x_i|y=1)}{f(x_i|y=0)}\right).$$

Hence, the log-density ratio can be expressed as the sum of p log-density ratios, one for each explanatory variable. Consequently, under the assumption that the predictors are conditionally independent given the response variable, the problem reduces to studying each individual predictor using the methodology for the one predictor case.

In more general cases relationships among the predictors are present, so we need to take into account their joint distribution. A simple and useful result may be obtained under the hypothesis of multivariate normality. As Scrucca (2003), we assume $\boldsymbol{x}|(y=0) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, $\boldsymbol{x}|(y=1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are the $p \times 1$ mean vectors, and $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are

the $p \times p$ covariance matrices. It can be shown that the log-odds is equal to

$$h(\boldsymbol{x}) = log \left(\frac{\frac{|\boldsymbol{\Sigma_1^{-1}}|^{1/2}}{(2\pi)^{p/2}} \left[\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_1})^T \boldsymbol{\Sigma_1^{-1}} (\boldsymbol{x}-\boldsymbol{\mu_1}) \right) \right]}{\frac{|\boldsymbol{\Sigma_0^{-1}}|^{1/2}}{(2\pi)^{p/2}} \left[\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_0})^T \boldsymbol{\Sigma_0^{-1}} (\boldsymbol{x}-\boldsymbol{\mu_0}) \right) \right]} \right) \tag{3.1}$$

$$= \frac{1}{2} log \left(\frac{|\boldsymbol{\Sigma_0}|}{|\boldsymbol{\Sigma_1}|} \right) + \boldsymbol{x^T} \left(\boldsymbol{\Sigma_1^{-1}} \boldsymbol{\mu_1} - \boldsymbol{\Sigma_0^{-1}} \boldsymbol{\mu_0} \right) + \frac{1}{2} \boldsymbol{x^T} \left(\boldsymbol{\Sigma_0^{-1}} - \boldsymbol{\Sigma_1^{-1}} \right) \boldsymbol{x}.$$

Equation (3.1) tells us that in general the terms required are x_j, x_j^2, and $x_j x_k$ $(j, k = 1, \cdots, p, j \neq k)$. If the two covariance matrices are equal, then the crossproduct and quadratic terms are no longer needed. If the variables are uncorrelated, the covariance matrices have off-diagonal elements equal to zero, and we do not need the crossproduct terms, but we still need the linear and quadratic terms. Hence, this case turns out to be one in which conditional independence holds. In fact, under the assumption of multivariate normal distribution for the two subpopulations, if the covariance matrix is diagonal, which implies that the variables are uncorrelated, they are also independent. In consequence, we can study each single predictor separately and evaluate which quadratic terms are needed.

In practical applications, the hypothesis of multivariate normality must be checked by the methodology introduced by Velilla (1993), who proposed a generalization of the Box-Cox approach for a simultaneous transformation of a set of variables to normality. He assumes there exists a vector $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p)^T$ of transformation parameters such that when we transform each $x_j (j = 1, ..., p)$ the following model holds $\boldsymbol{x^{(\lambda)}} = (x_1^{(\lambda_1)}, ..., x_p^{(\lambda_P)})^T \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma})$ where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ and $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$. The estimates for the λ s are obtained numerically maximizing the corresponding log-likelihood profile and can be obtained by Arc (Cook and Weisberg, 1999), a statistical software written in the Xlisp-Stat programming language. Arc is available for free from the Web at http://www.stat.umn.edu/arc.

4. Example

We consider the data on the recumbent cows presented in Clark *et al.* (1987). For unknown reasons, many pregnant dairy cows become recumbent–they lie down–either shortly before or after calving. This condition can be serious, and may lead to death of the cow. These data are taken from a study of blood samples of over 500 cows done at the Ruakura Animal Health Laboratory in New Zealand during 1983-84. A variety of blood tests were performed, and for many of the animals the *outcome* (survived, died) was determined. The goal is to see if survival can be predicted from the 8 blood measurements. In this study, we use two of the measurements CK and $UREA$, where CK is serum creatine phosphokinase and $UREA$ is serum urea.

From the previous analysis, we know that both $log(CK)$ and $log(UREA)$ will likely be needed in any logistic model for the regression of outcome. By applying the methodology introduced by Velilla (1993), we found out that the two conditional distributions of $\boldsymbol{x}|(outcome = 0)$ and $\boldsymbol{x}|(outcome = 1)$ are approximately multivariate normal, where $\boldsymbol{x} = (log(CK), log(UREA))^T$. The two covariance matrices are not equal, while correla-

tions of $log(CK)$ and $log(UREA)$ are very small in both conditional distributions. Thus, the crossproduct terms are not needed, but we still need the linear and quadratic terms.

A reasonable first logistic regression would include the five terms; $log(CK)$, $log(UREA)$, $[log(CK)]^2$, $[log(UREA)]^2$, and $[log(CK)][log(UREA)]$. The number of cases is now reduced to 259; this is the number of cows in which all the terms and the response were observed. Table 4.1 shows that two quadratic terms are required, but the crossproduct term is not needed. The summary of the fit of the logistic regression model without the crossproduct terms is presented in Table 4.2. The change in deviance is 271.171-269.962 = 1.209 with 254-253=1 df. The p -value from the χ^2_1 distribution is larger than 0.27. As expected, only the linear and quadratic terms are required.

Table 4.1 Logistic regression summaries for the recumbent data with 5 terms

Parameter	Estimate	Std. Error	Est/SE	p-value
Constant	-1.27137	5.63513	-0.226	0.8215
$log(CK)$	0.501139	0.980565	0.511	0.6093
$log(UREA)$	2.77473	3.31717	0.836	0.4029
$[log(CK)]^2$	-0.129480	0.0582141	-2.224	0.0261
$[log(UREA)]^2$	-1.47029	0.649189	-2.265	0.0235
$[log(CK)][log(UREA)]$	0.330085	0.300063	1.100	0.2713
Number of cases:				435
Number of cases used:				259
Degrees of freedom:				253
Pearson X2:				282.479
Deviance:				269.962

Table 4.2 Logistic regression summaries for the recumbent data with 4 terms

Parameter	Estimate	Std. Error	Est/SE	p-value
Constant	-6.13587	3.76992	-1.628	0.1036
$log(CK)$	1.20107	0.785803	1.528	0.1264
$log(UREA)$	5.18521	2.58269	2.008	0.0447
$[log(CK)]^2$	-0.131940	0.0591348	-2.231	0.0257
$[log(UREA)]^2$	-1.49602	0.636688	-2.350	0.0188
Number of cases				435
Number of cases used:				259
Degrees of freedom:				254
Pearson X2:				380.798
Deviance:				271.171

5. Remarks

In this article we considered logistic regression models for binary responses. The log-density ratio can be quite helpful for guiding the model development. Relevant statistical information can be extracted investigating the inverse problem, that is the distribution of the predictors given the response variable.

When we assume that the conditional distributions are multivariate normal, the linear, quadratic and crossproduct terms are required in general. If the two covariance matrices are equal, then the crossproduct and the quadratic terms are no longer needed. If the variables

are uncorrelated, we do not need the crossproduct terms. When our assumption does not hold, the log-density ratio may be derived under different parametric assumption.

References

Anderson, J. A. (1984). Regression and ordered categorical variables (with discussion). *Journal of Royal Statistical Society*, B, **46**, 1-30.

Clark, R. G., Henderson, H. V., Hoggard, G. K. Ellison, R. S., and Young, B. J. (1987). The ability of biochemical and haematological tests to predict recovery in periparturient recumbent cows. *New Zealand Veterinary Journal,* **35**, 126-133.

Cook, R. D. and Weisberg, S. (1999). *Applied regression including computing and Graphics,* John Wiley & Sons, New York.

Cox, D. R. (1970). *Analysis of binary data,* Chapman Hall, London.

Kay, R. and Little, S. (1987), Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, **74**, 495-501.

Kahng, M. (2005). Exploring interaction in generalized linear models. *Journal of Korean Data & Information Science Society,* **16**, 13-18.

Kahng, M. and Kim, M. (2004). A score test for detection of outliers in generalized linear models. *Journal of Korean Data & Information Science Society,* **15**, 129-139.

Kahng, M. and Kim, B. and Hong, J. (2010). Graphical regression and model assessment in logistic model. *Journal of Korean Data & Information Science Society,* **21**, 21-32.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of Royal Statistical Society*, B, **42**, 109-142.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, 2nd Ed., Chapman Hall, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of Royal Statistical Society*, A, **135**, 370-384.

Scrucca, L. (2003). Graphics for studying logistics regression models, *Statistical Methods and Applications,* **11**, 371-394.

Scrucca, L. and Weisberg, S. (2004), A simulation study to investigate the behavior of the log-density ratio under normality, *Communication in Statistics Simulation and Computation,* **33**, 159-178.

Seo, M. and Kim, J. (2006). Estimation of odds ratio in proportional odds model. *Journal of Korean Data & Information Science Society*, **17**, 1067-1076.

Velilla, S. (1993). A note on the multivariate Box-Cox transformations to normality. *Statistics and Probability Letters*, **17**, 315-322.