▨연구논문

The Confidence Regions for the Logistic Response Surface Model

Tae-Kyoung Cho
Division of Computer & Information Science, Dongguk University

Abstract

In this paper I discuss a method of constructing the confidence region for the logistic response surface model. The construction involves application of a general fitting procedure because the log odds is linear in its parameters. Estimation of parameters of the logistic response surface model can be accomplished by maximum likelihood, although this requires iterative computational method. Using the asymptotic results, asymptotic covariance of the estimators can be obtained. This can be used in the construction of confidence regions for the parameters and for the logistic response surface model.

I. Introduction

There have been lots of papers for the response surface analysis since Box and Wilson's study(1951). But it was not easy to find papers that the responses are binary variables in the response surface models.

In recent, the logistic model with the binary responses has been used in a variety of application. Brand et al.(1973) discussed the confidence bands for logistic response curve with one independent variable. The confidence bands for a logistic regression model with more than one independent variable was considered by Hauk(1983). Carter et al.(1986) discussed a conservative confidence region for the logistic response surface model.

When the binary response variable is affected by independence variables, the logistic model is often employed. Let Y denote a binary response variable that takes on values of 1 or 0 and $\mathbf{x}$ denote independent variables. For instance, Y

might indicate diagnosis of breast cancer (present, absent) or choice of automobile (domestic, foreign import). Let the probability of response surface given $\mathbf{x}$ be $\Pr(Y=1|\mathbf{x})=\pi(\mathbf{x})$ to be simplified notation. The logistic response surface model will be written as

$$\pi(\mathbf{x})=[1+\exp(-(\beta_0+\mathbf{x}'\boldsymbol{\beta}+\mathbf{x}'\mathbf{B}\mathbf{x}))]^{-1} \tag{1}$$

where

$$\mathbf{x}=[x_1\ x_2\ \cdots\ x_n]',$$

$$\boldsymbol{\beta}=[\beta_1\ \beta_2\ \cdots\ \beta_n]',$$

and

$$\mathbf{B}=\begin{bmatrix}\beta_{11} & \beta_{12}/2 & \cdots & \beta_{1n}/2\\ & \beta_{22} & \cdots & \beta_{2n}/2\\ & & \ddots & \vdots\\ & \text{symmetric} & & \beta_{nn}\end{bmatrix}.$$

To make a model (1) for a quadratic response surface and estimation easer, take the natural logarithm of odds (also known as the logit model), producing:

$$\ln[\pi(\mathbf{x})/(1-\pi(\mathbf{x}))]=\beta_0+\mathbf{x}'\boldsymbol{\beta}+\mathbf{x}'\mathbf{B}\mathbf{x} \tag{2}$$

Let $\mathbf{b}$ and $\mathbf{z}$ be vectors with $(1+n^*)$, $n^*=n(n+3)/2$, components such as

$$\mathbf{b}=[b_0,\ b_1,\ b_2,\ \cdots,\ b_{n^*}]'$$

$$=[\beta_0 \vdots \beta_1\cdots\beta_n \vdots \beta_{11}\beta_{12}\cdots\beta_{1n} \vdots \beta_{22}\beta_{23}\cdots\beta_{2n} \vdots \cdots \vdots \beta_{nn}]'$$

and

$$\mathbf{z}=[z_0,\ z_1,\ z_2,\ \cdots,\ z_{n^*}]'$$

$$=[1 \vdots x_1\cdots x_n \vdots x_1^2\ x_1x_2\cdots x_1x_n \vdots x_2^2\ x_2x_3\cdots x_2x_n \vdots \cdots \vdots x_n^2]'$$

Then the model (1) can be written as

$$\pi(\mathbf{x})=\pi(\mathbf{z})=(1+\exp(-\mathbf{z}'\mathbf{b}))^{-1} \tag{3}$$

It is clear that the logistic response surface model can be fitted in the same way as the logistic regression model and with the same computer programs.

It is assumed that K binary responses are independent Bernoulli random variables. When more than one observation, $n_i > 1$, on Y_i occur at a fixed $\mathbf{x_i} = [x_{i1}\ x_{i2}\ \cdots\ x_{in}]'$, the Y_i are independent binomial variables with $E(Y_i) = n_i\,\pi(\mathbf{z_i})$, $i = 1,2,\cdots,K$. The log likelihood function is

$$L(\mathbf{b}) = \sum_{i=1}^{K}\left[\ln\binom{n_i}{y_i} + y_i\,\mathbf{z_i}'\,\mathbf{b} - n_i \ln(1+\exp(\mathbf{z_i}'\,\mathbf{b}))\right] \tag{4}$$

where

$$\mathbf{z_i} = [z_{i0},\ z_{i1},\ z_{i2},\ \cdots,\ z_{in^*}]'$$
$$= [1 \vdots x_{i1}\cdots x_{in} \vdots x_{i1}^2\ \ x_{i1}x_{i2}\cdots x_{i1}x_{in} \vdots x_{i2}^2\ \ x_{i2}x_{i3}\cdots x_{i2}x_{in} \vdots \cdots \vdots x_{in}^2]'.$$

The score vector is

$$\partial L(\mathbf{b})/\partial\,\mathbf{b} = \sum_{i=1}^{K}[y_i - n_i\,\pi(\mathbf{z_i})]\mathbf{z_i}$$

The likelihood equations are obtained by setting $\partial L(\mathbf{b})/\partial\,\mathbf{b}$ to the zero vector. We can use the Newton-Raphson method to solve the likelihood equations. For details on the Newton-Raphson method, see Bard (1974). The observed information matrix is

$$\begin{aligned}\mathbf{J} &= -\partial^2 L(\mathbf{b})/\partial\,\mathbf{b}\partial\,\mathbf{b}' \\ &= \mathbf{Z}'\,diag[n_i\,\pi(\mathbf{z_i})(1-\pi(\mathbf{z_i}))]\,\mathbf{Z}\end{aligned} \tag{5}$$

where

$\mathbf{Z}' = [\ \mathbf{z}_0 \vdots\ \mathbf{z}_1 \vdots\ \mathbf{z}_2 \vdots\ \cdots\ \vdots\ \mathbf{z}_K\ ]$ is an $K \times n^*$ matrix and

$diag[n_i\pi(\mathbf{z_i})(1-\pi(\mathbf{z_i}))]$ denotes an $n^* \times n^*$ diagonal matrix.

Since the second partial derivatives given in (5) are not a function of $\{Y_i\}$, the observed and the expected information matrix are identical. Bradley (1962) discussed asymptotic normality of maximum likelihood estimators. For a large

sample size N, $N=\sum_{i=1}^{K} n_i$, maximum likelihood estimates, $\hat{\mathbf{b}}$, are asymptotically normal. The asymptotic result is

$$\sqrt{N}(\hat{\mathbf{b}}-\mathbf{b}) \xrightarrow[N\to\infty]{d} N_{n^*}(\mathbf{0},\ \boldsymbol{\Sigma})$$

where

$\xrightarrow[N\to\infty]{d}$ denotes convergence in distribution and

$N_{n^*}(\mathbf{0}, \boldsymbol{\Sigma})$ denotes n^*-dimensional multivariate normal distribution with mean zero vector $\mathbf{0}$ and the asymptotic covariance matrix $\boldsymbol{\Sigma}$.

The asymptotic covariance matrix for $\hat{\mathbf{b}}$ is estimated from the inverse of the sample information matrix and the estimated asymptotic covariance matrix is given by

$$\hat{\boldsymbol{\Sigma}}/N = \hat{\mathbf{J}}^{-1} = \{\ \mathbf{Z}'\, diag[n_i \hat{\pi}(\mathbf{z}_i)(1-\hat{\pi}(\mathbf{z}_i))]\ \mathbf{Z}\}^{-1}$$

where $\hat{\pi}(\mathbf{z}_i) = [1+\exp(-\mathbf{z}_i'\hat{\mathbf{b}})]^{-1}$

II. The Large Sample Confidence Regions for the Logistic Response Surface Model

I consider the problem of constructing the confidence region for logistic response surface model using a confidence set of parameters. These constructions are simplified by applying the logit transformation. I define the notations :

$\mathbf{D} = diag(\lambda_k)$ is the diagonal matrix of the eigenvalues of $\hat{\mathbf{J}}$.

$\mathbf{U}$ is the matrix of corresponding orthogonal eigenvectors.

$\mathbf{D}^{1/2} = diag(\lambda_k^{1/2})$.

$\mathbf{d}_i = [d_{i0},\ d_{i1},\ \cdots,\ d_{in^*}]' = \mathbf{D}^{1/2}\mathbf{U}'\mathbf{z}_i$.

$\boldsymbol{\eta} = [\eta_0,\ \eta_1,\ \cdots,\ \eta_{n^*}]' = \mathbf{D}^{1/2}\mathbf{U}'\mathbf{b}$.

$$\hat{\boldsymbol{\eta}} = [\ \hat{\eta}_0,\ \ \hat{\eta}_1,\ \cdots,\ \ \hat{\eta}_{n^*}]' = \mathbf{D}^{1/2}\ \mathbf{U}'\ \hat{\mathbf{b}}.$$

$$\boldsymbol{\theta} = [\ \theta_0,\ \ \theta_1,\ \cdots,\ \ \theta_{n^*}]' = \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}.$$

Then I have the multivariate standard normal distribution such that

$$\boldsymbol{\theta} \sim N_{n^*}(\mathbf{0},\ \mathbf{I})$$

where $\mathbf{I}$ is an $n^* \times n^*$ identity matrix.

Since the components of $\boldsymbol{\theta}$ are independent, I have

$$\begin{aligned} &\Pr\{-c_{\alpha/2} \le \theta_0 \le c_{\alpha/2},\ \ -c_{\alpha/2} \le \theta_1 \le c_{\alpha/2},\ \cdots,\ \ -c_{\alpha/2} \le \theta_{n^*} \le c_{\alpha/2}\} \\ &= \Pr\{-c_{\alpha/2} \le \theta_0 \le c_{\alpha/2}\}\Pr\{-c_{\alpha/2} \le \theta_1 \le c_{\alpha/2}\}\ \cdots\ \Pr\{-c_{\alpha/2} \le \theta_{n^*} \le c_{\alpha/2}\} \qquad (6) \\ &= 1-\alpha \end{aligned}$$

where $c_{\alpha/2}$ is a number such that

$$\int_{-c_{\alpha/2}}^{c_{\alpha/2}} \frac{1}{\sqrt{2\pi}} \exp(-\theta_j^2/2)\ d\theta_j = (1-\alpha)^{1/(n^*+1)},\ \ j = 0, 1, \cdots, n^*.$$

Substituting $\theta_j = \hat{\eta}_j - \eta_j$ into (6), we have the following a rectangular confidence set on $\boldsymbol{\eta}$ with confidence coefficient of $1-\alpha$:

$$\begin{cases} \hat{\eta}_0 - c_{\alpha/2} \le \eta_0 \le \hat{\eta}_0 + c_{\alpha/2} \\ \hat{\eta}_1 - c_{\alpha/2} \le \eta_1 \le \hat{\eta}_1 + c_{\alpha/2} \\ \qquad\qquad \vdots \\ \hat{\eta}_{n^*} - c_{\alpha/2} \le \eta_{n^*} \le \hat{\eta}_{n^*} + c_{\alpha/2} \end{cases} \qquad (7)$$

From a confidence set on $\boldsymbol{\eta}$ in (7), we have an inequality

$$\mathbf{d_i}'\hat{\boldsymbol{\eta}} - c_{\alpha/2}\left(\sum_{j=0}^{n^*} |d_{ij}|\right) \le \mathbf{d_i}'\boldsymbol{\eta} \le \mathbf{d_i}'\hat{\boldsymbol{\eta}} + c_{\alpha/2}\left(\sum_{j=0}^{n^*} |d_{ij}|\right) \text{ for all } \mathbf{d_i} \qquad (8)$$

This is equivalent to the inequality of (9)

$$\mathbf{z}_i'\hat{\mathbf{b}} - c_{a/2}\left(\sum_{j=0}^{n} |d_{ij}|\right) \leq \mathbf{z}_i'\mathbf{b} \leq \mathbf{z}_i'\hat{\mathbf{b}} + c_{a/2}\left(\sum_{j=0}^{n} |d_{ij}|\right) \text{ for all } \mathbf{z}_i \qquad (9)$$

Therefore, 100(1-a)% confidence region for $\mathbf{z}_i'\mathbf{b}$ over all $\mathbf{z}_i$ are given by

$$[L_b, U_b] = \left[\ \mathbf{z}_i'\hat{\mathbf{b}} - c_{a/2}\left(\sum_{j=0}^{n} |d_{ij}|\right),\quad \mathbf{z}_i'\hat{\mathbf{b}} + c_{a/2}\left(\sum_{j=0}^{n} |d_{ij}|\right)\ \right]$$

Since $\pi(\mathbf{z}_i) = \pi(\mathbf{x}_i)$, $i=1, \cdots, K$, the corresponding 100(1-a)% confidence region on $\pi(\mathbf{x}_i)$ over all $\mathbf{x}_i$ are given by taking the inverse logit transform of inequality in (9) :

$$[1+\exp(-L_b)]^{-1} \leq \pi(\mathbf{x}_i) \leq [1+\exp(-U_b)]^{-1}$$

Ⅲ. Numerical Example

I shall illustrate the procedure with the example, shown in <Table 1>, which is provided by Carter et al.(1986). <Table 1> lists a combination of methylmethanesulfonate (MMS) and phorbol 12-myristate, 13-acetate (PMA) in the human promyelocytic leukemia cell line HL-60. Based on the dose-response curve for each individual agent, it was of interest to evaluate the dose-response surface for their combination. I consider the following model for such curvature in the dose-response relationship.

$$\pi(\mathbf{x}_i) = [1+\exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2))]^{-1} \qquad (10)$$

where

$\pi(\mathbf{x}_i)$ is proportion of dead cells,

x_1 = concentration of MMS and

x_2 = concentration of PMA.

Since the p-value is 0.7053 for test of $H_0 : \beta_{12} = 0$, it was deleted from the

model in (10). Using SAS/IML (1990), we reanalyze the data and have the maximum likelihood estimates of $\boldsymbol{b}=(\beta_0,\ \beta_1,\ \beta_2,\ \beta_{11},\ \beta_{22})'$ and the estimated asymptotic covariance matrix of $\hat{\boldsymbol{b}}$ such as :

$$\hat{\boldsymbol{b}}=(-1.3299,\ -0.00835,\ 0.1591,\ 0.000039,\ -0.00131)'$$

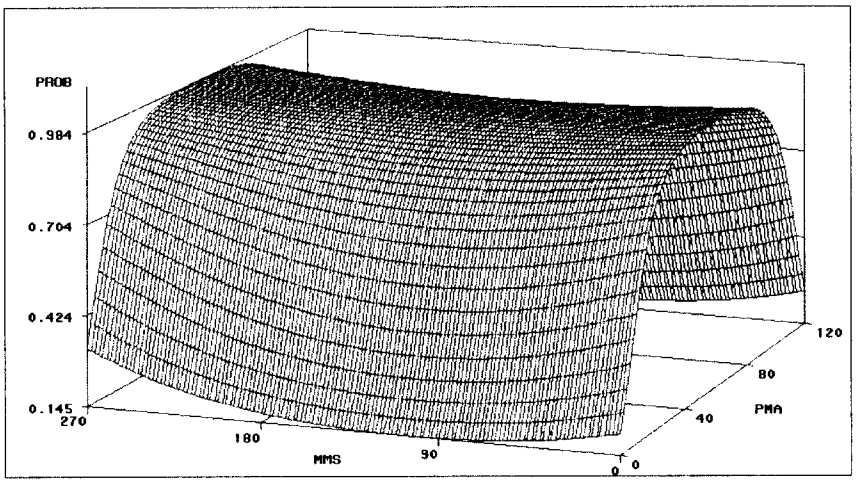
and

$$Cov(\hat{\boldsymbol{b}})=\begin{bmatrix} 1.389\times10^{-2} & -1.280\times10^{-4} & -1.079\times10^{-3} & 3.679\times10^{-7} & 9.878\times10^{-6} \\ & 7.083\times10^{-6} & -2.017\times10^{-6} & -2.673\times10^{-8} & 1.613\times10^{-8} \\ & & 2.672\times10^{-4} & 9.441\times10^{-9} & -2.565\times10^{-6} \\ & & & 1.063\times10^{-10} & -7.580\times10^{-11} \\ \text{symmetric} & & & & 2.485\times10^{-8} \end{bmatrix}$$

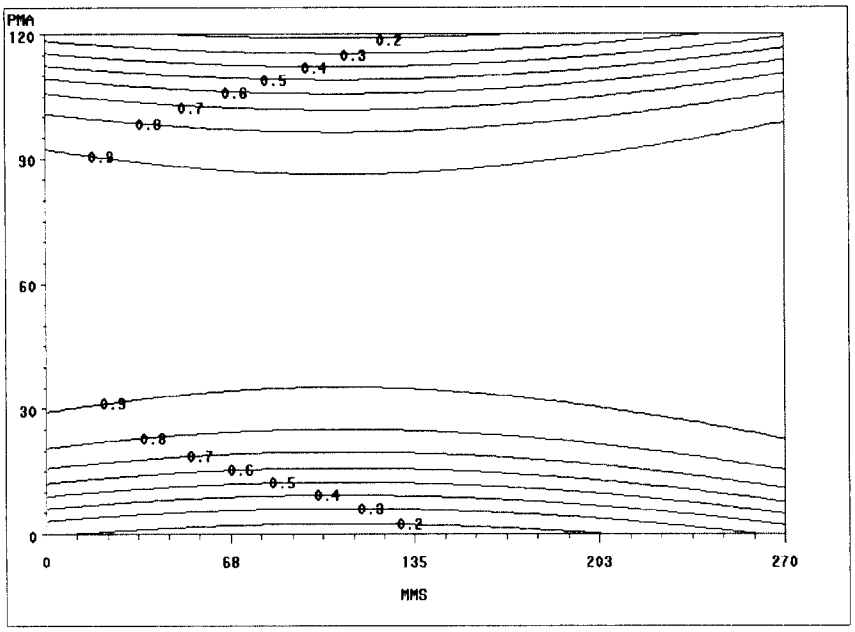
The fitted logistic response surface and the contours of the fitted response surface are given in <Figure 1> and <Figure 2>, respectively. <Table 2> shows the expected number of dead cells and 95% confidence regions of fitted proportion of dead cells for fixed levels of PMA and MMS.

< Table 1 > Treatment Combinations and Observation

MMS (μg/ml)	PMA ($M\times10^{-9}$)	No. viable cells	No. dead cells
0	0	79	19
0	1	63	24
0	10	37	54
0	100	19	68
10	0	75	16
100	0	73	17
250	0	73	19
10	1	75	19
10	10	49	41
10	100	14	79
100	1	73	10
100	10	49	36
100	100	21	62
250	1	56	36
250	10	37	56
250	100	13	74



< Figure 1 > The fitted logistic response surface



< Figure 2 > The contours of the < Figure 1 >

< Table 2 > The 95% confidence regions of the fitted proportions of dead cells and the expected number of dead cells

MMS (μg/ml)	PMA ($M \times 10^{-9}$)	Lower bound	Fitted proportions of dead cells	Upper bound	Expected no. of dead cells
0	0	0.162	0.209	0.266	20.5
0	1	0.183	0.236	0.300	20.6
0	10	0.429	0.533	0.634	48.5
0	100	0.699	0.819	0.898	71.2
10	0	0.146	0.196	0.259	17.9
100	0	0.079	0.145	0.251	13.0
250	0	0.164	0.271	0.412	24.9
10	1	0.170	0.222	0.285	20.9
10	10	0.408	0.513	0.616	46.1
10	100	0.705	0.806	0.879	75.0
100	1	0.096	0.165	0.270	13.7
100	10	0.270	0.422	0.591	35.8
100	100	0.557	0.743	0.869	61.6
250	1	0.196	0.303	0.436	27.9
250	10	0.439	0.615	0.766	57.2
250	100	0.721	0.864	0.939	75.1

References

[1] Bard, Y.(1974), *Nonlinear Parameter Estimation*, Academic Press, New York

[2] Box, G.E.P. and Wilson, K.B.(1951), "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society*, B-13, pp. 1-38.

[3] Bradley, R.A. and Gart, J.J.(1962), "The Asymptotic Properties of ML Estimations when Sampling from Associated Populations," *Biometrika*, Vol. 49, pp. 205-214.

[4] Brand, R.J., Pinnock, D.E. and Jackson, K.L.(1973), "Large Sample Confidence Bands for the Logistic Response Curve and Its Inverse," *The American Statistician*, Vol. 27, pp. 157-160

[5] Carter, Jr. W.H., Chinchilli, V.M., Wilson, J.D., Cambell, E.D., Kessler, F.K. and Carchman, R.A.(1986), "An Asymptotic Confidence Region for the ED_{100p} from the Logistic Response Surface for a Combination of Agents," *The American Statistician*, Vol. 40, pp.124-128

[6] Hauck, W.W.(1983), "A Note on Confidence Bands for the Logistic Response Curve," *The American Statistician*, Vol, 37, pp. 158-160.

[7] SAS Institute Inc.(1990), *SAS/IML*