Far East Journal of Theoretical Statistics
Volume 39, Number 1, 2012, Pages 53-66
Published Online: June 2012
Available online at http://pphmj.com/journals/fjts.htm
Published by Pushpa Publishing House, Allahabad, INDIA

INFERENCES IN LOGISTIC MODELS ALONG WITH ESTIMATION OF PARAMETERS USING DISCRIMINANT FUNCTION APPROACH

Mezbahur Rahman, Katherine A. Holt and Patrick A. O'Connor

Minnesota State University
Mankato, U. S. A.
e-mail: mezbahur.rahman@mnsu.edu

Abstract

Parameter estimation in logistic regression using the maximum likelihood approach and the discriminant function approach is revisited. A bootstrap estimation of the standard errors for the estimates of the model parameters in discriminant function approach is obtained. The asymptotic distributions are compared in both the estimation procedures using bootstrap sampling. All the computations are performed for three different vastly used data sets and analyzed.

1. Introduction

In a binary response model, it cannot be assumed that the errors have a normal distribution and hence usual linear regression is not applicable. Due to a wide range of applications, the binary response models are studied explicitly. Here we discuss the logistic regression model. For the latest developments in the area, the reader is referred to Cox and Snell [1],

2010 Mathematics Subject Classification: 62Fxx.

Keywords and phrases: binary response model, maximum likelihood estimate, least square estimate, Newton-Raphson method.

Received December 15, 2011

McCullagh and Nelder [3], Ryan [6], Hosmer and Lemeshow [2], Powers and Xie [4], and the references therein.

Initial notation for the model is now as follows: Let $X_1, X_2, ..., X_p$ be p different regressors and Y be the response (dependent) variable. Y can only take the values of '1' for 'success' and '0' for 'failure'. A random sample of n data points is taken from a phenomenon. A general binary model is assumed as

$$P(Y_i = 1 | X_{1i}, X_{2i}, ..., X_{pi}) = \Lambda_i = E(Y_i | X_{1i}, X_{2i}, ..., X_{pi}), \quad i = 1, 2, ..., n,$$

where $\Lambda_i \in [0, 1]$ and $P(Y_i = 0) = 1 - \Lambda_i$. We define the logistic regression model as

$$\Lambda(\mathbf{x}_i'\beta) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}, \tag{1}$$

where $\beta_0, \beta_1, ..., \beta_p$ are unknown constants. Notice that there is no error term on the right side of (1), because the left side is a function of $E(Y | X_1, X_2, ..., X_p)$, instead of Y, which serves to remove the error term.

2. Motivation

In logistic regression, maximum likelihood estimation procedure became only viable procedure as almost all the commonly used softwares use the method. As due to nature of data, often the design matrices are ill behaved but with the help of the generalized inverses of the matrices that difficulty can be overcome. But the inferences about the parameters are less reliable as the underlying asymptotic distributions might be affected by the multicollinearity among the regressors. Hence the initiation is taken in this paper to compare bootstrap distributions of the respective statistics in maximum likelihood estimation procedure and in discriminant function approach.

In the following sections, we describe the commonly used maximum likelihood estimation procedure and inferences about the parameters, and the

discriminant function approach in estimating parameters and their inferences. Then we apply the procedures for three different vastly used data sets. Finally, we give a brief conclusion about the findings in the data sets used and some general comments.

3. Maximum Likelihood Estimation (MLE)

From (1), we have

$$\ln\left[\frac{\Lambda(\mathbf{x}'\beta)}{1-\Lambda(\mathbf{x}'\beta)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}'\beta,$$

where $\mathbf{x}' = [1, x_1, x_2, ..., x_p]$ and 'ln' stands for the natural logarithm. The estimators are generally obtained by maximizing the logarithm of the likelihood function. The likelihood on data with *n* binary responses may be written as

$$L = \prod_{i=1}^{n} [\Lambda(\mathbf{x}_i'\beta)]^{y_i} [1-\Lambda(\mathbf{x}_i'\beta)]^{1-y_i},$$

where $\Lambda(\mathbf{x}_i'\beta)$ is defined in (1). The log-likelihood function (log stands for natural logarithm) is

$$\ln L = \sum_{i=1}^{n} \{y_i \ln[\Lambda(\mathbf{x}_i'\beta)] + (1-y_i)\ln[1-\Lambda(\mathbf{x}_i'\beta)]\}. \quad (2)$$

Because $\Lambda(\mathbf{x}_i'\beta)$ is nonlinear in the unknown parameters, we solve the likelihood equations derived from (2) iteratively using the Newton-Raphson method. The first and the second derivatives, which are used to maximize the log-likelihood, are given in the following expressions:

$$\frac{\delta \ln(L(\beta))}{\delta\beta} = \mathbf{U}(\beta) = \sum_{i=1}^{n} [y_i - \Lambda(\mathbf{x}_i'\beta)]\mathbf{x_i}, \text{ and}$$

$$-\left[\frac{\delta^2 \ln(L(\beta))}{\delta\beta\beta'}\right] = \mathbf{I}(\beta) = \sum_{i=1}^{n} \Lambda(\mathbf{x}_i'\beta)\mathbf{x_i}\mathbf{x}_i'. \quad (3)$$

At the *t*th iteration, the estimates are obtained using the equation

$$\hat{\beta}^{(t)} = \hat{\beta}^{(t-1)} + [\mathbf{I}(\hat{\beta}^{(t-1)})]^{-1}\mathbf{U}((\hat{\beta})^{(t-1)}),$$

where $\hat{\beta}^{(0)}$ is obtained by regressing y on the x's using the usual least squares method. The iteration is stopped when the consecutive iteration values are close and/or the log-likelihood values are maximized (see Powers and Xie [4] for details).

Significance of the individual parameter can be tested by assuming that the samples are large, using the test statistic

$$Z_{Lj} = \frac{\hat{\beta}_{Lj}}{s_{\hat{\beta}_{Lj}}}, \quad j = 0, 1, 2, ..., p, \tag{4}$$

where L stands for MLE, $s_{\hat{\beta}_{Lj}} = \sqrt{[\mathbf{I}(\hat{\beta})]^{-1}_{jj}}$, the square root of the jth diagonal element of the inverse of $\mathbf{I}(\hat{\beta})$ in (3) evaluated at $\hat{\beta}$. Then for a large sample, Z_{Lj} will have an approximate standard normal distribution under the null hypothesis, $H_0 : \beta_j = 0$.

For comparison, here we will also use bootstrap estimate of the standard error and (4) can be written as

$$Z_{Bj} = \frac{\hat{\beta}_{Lj}}{s_{\hat{\beta}_{Bj}}}, \quad j = 0, 1, 2, ..., p, \tag{5}$$

where B stands for bootstrap. Asymptotic distribution of Z_{Bj} can be assumed the standard normal and hence the p-values can be obtained.

On the other hand, the respective p-values for the significances of the variables can also be computed using the usual bootstrap method.

4. Discriminant Function Approach (DFA)

In a dichotomous response model, discriminant function approach is a

well established parameter estimation procedure. The discriminant function approach in estimating parameters in the logistic regression models is based on the assumption that the distribution of the regressors, given the value of the response, is multivariate normal. Even though, this assumption is rarely valid, the estimates are plausible and simplistic in nature as they are based on sample means and covariances.

The conditional distribution of the regressors (X) given the response $(y = 1, 0)$ is the following multivariate normal distributions:

$$\mathbf{X} \mid \mathbf{y} = \mathbf{k} \sim \mathbf{N}(\mu_{\mathbf{k}}, \boldsymbol{\Sigma}_{\mathbf{k}}),$$

where $k = 1$ (success), 0 (failure), and $\mu_{\mathbf{k}}$ contains respective means and $\boldsymbol{\Sigma}_{\mathbf{k}}$ contains respective covariances for p regressors. Under these assumptions, the intercept coefficient in the logistic regression can be written as (see Hosmer and Lemeshow [2, p. 43]):

$$\beta_0 = \ln\left(\frac{\theta_1}{\theta_0}\right) - 0.5(\mu_1 - \mu_0)' \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_0) \tag{6}$$

and the $p \times 1$ vector of the slope parameters can be written as

$$\beta = (\mu_1 - \mu_0)' \boldsymbol{\Sigma}^{-1}, \tag{7}$$

where $\theta_1 = P(Y = 1)$ and $\theta_0 = P(Y = 0) = 1 - \theta_1$ denote the respective probabilities.

The discriminant function estimators of β_0 and β are found by substituting the respective sample estimates $\tilde{\mu}_0 = \overline{\mathbf{X}}_0$ and $\tilde{\mu}_1 = \overline{\mathbf{X}}_1$, the sample mean vectors for the regressors at $y = 0$ and $y = 1$, and $\tilde{\boldsymbol{\Sigma}} = \mathbf{S} = \dfrac{(\mathbf{n_0} - \mathbf{1})\mathbf{S_0} + (\mathbf{n_1} - \mathbf{1})\mathbf{S_1}}{(\mathbf{n_0} + \mathbf{n_1} - \mathbf{2})}$, where $\mathbf{S_0}$ and $\mathbf{S_1}$ are the sample variance covariance $p \times p$ matrices for the regressors at $y = 0$ and $y = 1$, respectively, and n_0 and n_1 are the numbers of 0's and 1's for the response variable. And the estimates of θ_0 and θ_1 are the respective sample

proportions $\tilde{\theta}_0 = \frac{n_0}{n_0 + n_1}$ and $\tilde{\theta}_1 = \frac{n_1}{n_0 + n_1}$. Then $\hat{\beta}_{D0}$ and $\hat{\beta}_D$ $(p \times 1 \text{ vector})$ (*D* stands for DFA) can be computed after substituting the respective estimates in (6) and (7).

Inferences can be made using the bootstrap distribution and by computing test statistics similar to (5) as

$$Z_{Dj} = \frac{\hat{\beta}_{Dj}}{s_{\hat{\beta}_{Bj}}}, \quad j = 0, 1, 2, ..., p \tag{8}$$

and *B* stands for bootstrap. Asymptotic distribution of Z_{Dj} can be assumed the standard normal and hence the *p*-values can be obtained.

Due to the availability of relevant software and established asymptotic properties, the maximum likelihood estimates are commonly used. Here, we have compared the bootstrap distributions for the above mentioned estimators and showed that how effectively inferences can also be drawn in the discriminant function approach using the respective bootstrap distributions.

5. The Low Birth Weight Study (LBW) Data

Data were collected as a part of a larger study at Baystate Medical Center in Springfield, Massachusetts and included in Hosmer and Lemeshow [2] and can be accessed from http://www-unix.oit.umass.edu/~statdata. Code sheet for the variables in The Low Birth Weight Data is as follows:

Variable	Description	Codes/Values	Name
1	Identification Code	ID Number	ID
2	Low Birth Weight	0 = ≥ 2500g 1 = < 2500g	LOW
3	Age of Mother	Years	AGE
4	Weight of Mother at Last Menstrual Period	Pounds	LWT

5	Race	1 = *White* 2 = *Black* 3 = *Other*	RACE
6	Smoking Status During Pregnancy	0 = *No* 1 = *Yes*	SMOKE
7	History of Premature Labor	0 = *None* 1 = *One* 2 = *Two, etc*.	PTL
8	History of Hypertension	0 = *No* 1 = *Yes*	HT
9	Presence of Uterine Irritability	0 = *No* 1 = *Yes*	UI
10	Number of Physician Visits During the First Trimester	0 = *None* 1 = *One* 2 = *Two, etc*.	FTV
11	Birth Weight	Grams	BWT

The Race variable is of nominal scale of three distinct values and hence created two independent variables called *RACE*2 and *RACE*3 for Black and Other, respectively. Then all the coefficients are estimated using the both MLE and DFA methods. In odd numbered tables, MLE computations are displayed. And in even numbered tables, DFA computations are displayed. The rows of the tables are indicating the variables as defined in the description of the data. The columns for odd numbered tables are defined as follows:

$\hat{\beta}_{Lj}$	Maximum Likelihood Estimate
$s_{\hat{\beta}_{Lj}}$	Standard Error Using Information Matrix
$p_{Z_{Lj}}$	*p*-value Using Z_{Lj}
$\bar{\hat{\beta}}_{Lj}$	Mean of MLE's for 10000 Bootstrap Samples
$s_{\hat{\beta}_{Bj}}$	Standard Deviation of MLE's for Bootstrap Samples
p_{BOOT}	Bootstrap *p*-value

$p_{Z_{Bj}}$	Asymptotic *p*-value for MLE Using Z_{Bj}
p-Chisq	*p*-value for Chi-square Goodness-of-fit test for Normality of BZ using 100 groups
D2	Deviance $= \sum_{i=1}^{10000} (U_i - \Phi_i)^2$, where $U_i = 0.0001, 0.0002, ..., 1$ and Φ_i are the standard normal distribution function values for the ordered studentized Z_{Bj} values

The columns for even numbered tables are defined as follows:

Table 1. LBW Data: Maximum Likelihood Estimates (MLE)

Variable	$\hat{\beta}_{Lj}$	$s_{\hat{\beta}_{Lj}}$	$p_{Z_{Lj}}$	$\hat{\beta}_{Bj}$	$s_{\hat{\beta}_{Bj}}$	p_{BOOT}	$p_{Z_{Bj}}$	*p*-Chisq	D2
Intercept	141.20	45.49	0.0019	135.89	22.48	0.0000	0.0000	0.0000	1.71
AGE	–0.15	0.32	0.6320	–0.16	0.32	0.6012	0.6277	0.0000	1.98
LWT	0.01	0.05	0.9200	–0.00	0.06	0.9122	0.9314	0.0000	8.39
RACE2	3.04	7.77	0.6952	3.93	3.91	0.4832	0.4418	0.0000	1.34
RACE3	4.13	6.38	0.5173	1.42	2.92	0.3317	0.1603	0.0000	1.19
SMOKE	3.87	6.97	0.5785	0.39	0.59	0.2819	0.0000	0.0000	2.17
PTL	9.51	37.19	0.7983	8.33	3.71	0.0128	0.0102	0.0000	5.71
HT	5.76	9.60	0.5484	3.32	7.23	0.4771	0.4270	0.0000	32.66
UI	–4.72	5.16	0.3604	–6.52	3.57	0.2213	0.1918	0.0000	3.38
FTV	0.06	1.81	0.9752	0.39	0.59	0.9280	0.9229	0.0000	2.17
BWT	–0.06	0.02	0.0015	–0.05	0.01	0.0000	0.0000	0.0000	6.35

$\hat{\beta}_{Dj}$	DFA Estimate of the Parameters
$\bar{\hat{\beta}}_{Bj}$	Mean of DFA's for 10000 Bootstrap Samples
$s_{\hat{\beta}_{Bj}}$	Standard Deviation of DFA's for Bootstrap Samples
p_{BOOT}	Bootstrap *p*-value
Z_{Dj}	$\frac{DFA}{SBDFA}$
$p_{Z_{Dj}}$	Asymptotic *p*-value for Z_{Dj}
p-Chisq	*p*-value for Chi-square Goodness-of-fit test for Normality of Z_{Dj} using 100 groups

D2	Deviance $= \sum_{i=1}^{10000} (U_i - \Phi_i)^2$, where $U_i = 0.0001, 0.0002, ..., 1$ and Φ_i are the standard normal distribution function values for the ordered studentized Z_{Dj} values.

Table 2. LBW Data: Discriminant Function Approach (DFA)

Variable	$\hat{\beta}_{Dj}$	$\hat{\beta}_{Bj}$	$s_{\hat{\beta}_{Bj}}$	p_{BOOT}	Z_{Dj}	$p_{Z_{Dj}}$	p-Chisq	D2
Intercept	18.70	20.08	3.54	0.0001	5.29	0.0000	0.0000	2.58
AGE	–0.07	–0.08	0.08	0.3551	–0.93	0.3540	0.0000	0.75
LWT	–0.01	–0.01	0.01	0.6680	–0.42	0.6766	0.0958	0.20
RACE2	–0.30	–0.33	1.01	0.7536	–0.30	0.7651	0.5512	0.28
RACE3	–0.44	–0.47	0.65	0.4882	–0.67	0.5005	0.0014	0.27
SMOKE	–0.20	–0.23	0.63	0.7365	–0.32	0.7482	0.0082	0.34
PTL	1.15	1.30	0.60	0.0609	1.92	0.0545	0.0000	4.53
HT	0.87	0.93	1.34	0.5090	0.65	0.5183	0.6502	0.16
UI	–1.28	–1.36	0.97	0.1772	–1.32	0.1858	0.0106	0.39
FTV	–0.01	–0.02	0.28	0.9721	–0.03	0.9732	0.0077	0.81
BWT	–0.01	–0.01	0.01	0.0000	–7.75	0.0000	0.0000	3.68

6. Lung Cancer and Bird Keeping (LCB) Data

To investigate whether bird keeping is a risk factor, researchers in Hague, Netherlands conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among patients who were registered with a general practice, age 65 or younger, and had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure. Data is obtained from Ramsey and Schafer [5, Display 20.2]. The description of the data is as follows:

LC = Lung Cancer (1 = lung cancer patients, 0 = controls).

FM = Sex $(1 = F, 0 = M)$.

SS = Socioeconomic status (1 = High, 0 = Low), determined by occupation of the household's principal wage earner.

BK = Indicator of bird keeping (caged birds in the home for more

than 6 consecutive months from 5 to 14 years before diagnosis (cases) or examination (controls).

AG = Age, in years.

YR = Years of smoking prior to diagnosis or examination.

CD = Average rate of smoking, in cigarettes per day.

Table 3. LCB Data: Maximum Likelihood Estimates (MLE)

Variable	$\hat{\beta}_{Lj}$	$s_{\hat{\beta}_{Lj}}$	$p_{Z_{Lj}}$	$\hat{\beta}_{Bj}$	$s_{\hat{\beta}_{Bj}}$	p_{BOOT}	$p_{Z_{Bj}}$	*p*-Chisq	D2
Intercept	–1.38	1.75	0.4321	–1.22	2.12	0.5000	0.5154	0.0817	0.45
FM	–0.56	0.53	0.2907	–0.62	0.58	0.3274	0.3362	0.2073	0.27
SS	0.11	0.47	0.8221	0.12	0.51	0.8341	0.8369	0.6160	0.15
BK	1.36	0.41	0.0009	1.44	0.46	0.0063	0.0030	0.0845	0.59
AG	–0.04	0.04	0.2625	–0.05	0.05	0.3823	0.4296	0.0000	7.04
YR	0.07	0.03	0.0059	0.09	0.04	0.0684	0.0589	0.0000	26.06
CD	0.03	0.03	0.3081	0.03	0.03	0.3869	0.4027	0.0008	0.82

Table 4. LCB Data: Discriminant Function Approach (DFA)

Variable	$\hat{\beta}_{Dj}$	$\hat{\beta}_{Bj}$	$s_{\hat{\beta}_{Bj}}$	p_{BOOT}	Z_{Dj}	$p_{Z_{Dj}}$	*p*-Chisq	D2
Intercept	–2.22	–2.27	1.80	0.2124	–1.24	0.2165	0.2474	0.13
FM	–0.50	–0.53	0.55	0.3594	–0.91	0.3630	0.4250	0.16
SS	0.16	0.18	0.46	0.7242	0.34	0.7330	0.0051	0.63
BK	1.48	1.57	0.52	0.0094	2.83	0.0047	0.0000	2.69
AG	–0.01	–0.02	0.03	0.6311	–0.48	0.6338	0.1256	0.20
YR	0.05	0.06	0.02	0.0081	2.79	0.0052	0.5448	0.21
CD	0.02	0.02	0.03	0.4064	0.81	0.4187	0.0002	0.75

7. Intensive Care Unit (ICU) Data

The data that was collected by the Baystate Medical Center in Springfield, Massachusetts was used. As mentioned earlier, the data consists of 200 observations which were part of a study on the survival of patients following admission to an adult intensive care unit (ICU). The goal was to develop a logistic model to predict the probability of survival to hospital discharge of patient and to gain a better understanding of the risk factors associated with ICU mortality.

Variable	Description	Codes/Values	Name
1	Identification Code	ID Number	ID
2	Vital Status	0 = *Lived*	STA
		1 = *Died*	
3	Age	Years	AGE
4	Sex	0 = *Male*	SEX
		1 = *Female*	
5	Race	1 = *White*	RACE
		2 = *Black*	
		3 = *Other*	
6	Service at ICU Admission	0 = *Medical*	*SER*
		1 = *Surgical*	
7	Cancer Part of Present Problem	0 = *No*	CAN
		1 = *Yes*	
8	History of Chronic Renal Failure	0 = *No*	CRN
		1 =*Yes*	
9	Infection Probable at ICU	0 = *No*	INF
	Admission	1 = *Yes*	
10	CPR Prior to ICU Admission	0 = *No*	CPR
	During the First Trimester	1 = *Yes*	
11	Systolic Blood Pressure at ICU Admission	mm Hg	SYS
12	Heart Rate at ICU Admission	Meats/min	HR
13	Previous Admission to an ICU	0 = *No*	PRE
	Within 6 Months	1 = *Yes*	
14	Type of Admission	0 = *Elective*	TYP
		1 = *Emergency*	
15	15 Long Bone, Multiple, Neck,	0 = *No*	FRA
	Single Area, or Hip Fracture	1 = *Yes*	
16	PO2 from Initial Blood Gases	0 = > 60	PO2
		1 = ≤ 60	
17	PH from Initial Blood Gases	0 = ≥ 7.25	PH
		1 = < 7.25	
18	PCO2 from Initial Blood Gases	0 = ≤ 45	PCO
		1 = > 45	

19	Bicarbonate from Initial Blood Gases	0 = ≥ 18	BIC
		1 = < 18	
20	Creatinine from Initial Blood Gases	0 = ≤ 2.0	CRE
		1 = > 2.0	
21	Level of Consciousness at ICU Admission	0 = *No Coma*	LOC
		1 = *Deep Stupor*	
		2 = *Coma*	

All computations are performed using MATLAB software and the programs are available from the author upon request.

Table 5. ICU Data: Maximum Likelihood Estimates (MLE)

Variable	$\hat{\beta}_{Lj}$	$s_{\hat{\beta}_{Lj}}$	$p_{Z_{Lj}}$	$\hat{\beta}_{Bj}$	$s_{\hat{\beta}_{Bj}}$	p_{BOOT}	$p_{Z_{Bj}}$	p-Chisq	D2
Intercept	–5.39	1.20	0.0070	–8.01	6.16	0.2237	0.3815	0.0000	67.24
AG	0.05	0.02	0.0029	0.07	0.07	0.1670	0.4479	0.0000	130.31
SYS	–0.01	0.01	0.1720	–0.02	0.02	0.4568	0.5528	0.0000	29.65
HRA	–0.00	0.01	0.7890	–0.01	0.02	0.8695	0.9002	0.0000	19.30
RACE2	–0.81	1.17	0.4895	–4.46	4.32	0.6970	0.8518	0.0000	91.47
RACE3	0.41	1.11	0.7145	–0.92	4.23	0.8011	0.9233	0.0000	129.50
SEX	–0.55	0.50	0.2702	–0.94	1.11	0.4344	0.6180	0.0000	73.87
SER	–0.64	0.59	0.2807	–0.83	1.40	0.4685	0.6489	0.0000	54.19
CAN	2.75	0.98	0.0049	4.41	3.98	0.2290	0.4890	0.0000	138.19
CRN	–0.11	0.78	0.8873	–0.28	3.30	0.9301	0.9734	0.0000	174.42
INF	–0.07	0.53	0.8889	–0.22	1.08	0.9185	0.9449	0.0000	35.75
CPR	0.93	0.99	0.0072	0.83	4.17	0.5605	0.8229	0.0000	163.11
PRE	0.98	0.64	0.1239	1.35	1.34	0.3035	0.4641	0.0000	41.96
TYP	2.67	1.00	0.0072	5.37	4.20	0.2891	0.5242	0.0000	143.11
FRA	1.26	1.01	0.2132	1.09	3.05	0.4667	0.6802	0.0000	69.63
PO2	0.28	0.86	0.7476	–0.67	3.47	0.8651	0.9368	0.0000	123.84
PH	2.38	1.23	0.0537	4.06	7.72	0.3316	0.7580	0.0000	254.84
PCO	–3.15	1.38	0.0225	–6.38	10.99	0.2919	0.7747	0.0000	312.01
BIC	–0.77	0.91	0.4021	–1.20	5.44	0.6755	0.8881	0.0000	238.42
CRE	0.14	1.07	0.8958	0.11	4.19	0.9339	0.9733	0.0000	150.49
LOC	2.72	0.76	0.0003	7.34	8.02	0.3032	0.7343	0.0000	262.11

8. Concluding Remarks

For all three data sets, bootstrap standard errors for DFA method are lower compared to that of MLE methods.

Bootstrap *p*-values (p_{BOOT}) matched with asymptotic *p*-values $(p_{Z_{Bj}})$ in MLE only exception is in case of SMOKE variable in LBW data.

Bootstrap *p*-values (p_{BOOT}) matched with asymptotic *p*-values $(p_{Z_{Dj}})$ in DFA in all data sets.

In MLE, asymptotic *p*-values contradicted with bootstrap *p*-values in several instances, such as, in LBW data for the variable PTL, in LCB data for the variable YR, and in ICU data for the variables AG, CAN, CPR, TYP, PH, PCO and LOC.

The *Chi-square p*-values indicate that the asymptotic distributions are not normal except for some variables in LCB data which might be due to simplicity of this data set compared to the other two data sets.

For all data sets, the deviance statistic (D2) is lower for DFA method compared to the MLE method showing that bootstrap distributions for the respective statistics are closer to normal.

Table 6. ICU Data: Discriminant Function Approach (DFA)

Variable	$\hat{\beta}_{Dj}$	$\hat{\beta}_{Bj}$	$s_{\hat{\beta}_{Bj}}$	p_{BOOT}	Z_{Dj}	$p_{Z_{Dj}}$	*p*-Chisq	D2
Intercept	–5.09	–5.85	2.41	0.0485	–2.11	0.0348	0.0000	1.82
AG	0.04	0.05	0.02	0.0250	2.38	0.0175	0.0068	0.43
SYS	–0.01	–0.01	0.01	0.4038	–0.82	0.4132	0.0077	0.81
HRA	–0.01	–0.01	0.01	0.7706	–0.27	0.7860	0.0000	1.21
RACE2	–0.96	–1.14	0.83	0.2426	–1.17	0.2429	0.0132	0.18
RACE3	0.43	0.45	1.20	0.6957	0.36	0.7183	0.0000	2.03
SEX	–0.53	–0.62	0.59	0.3561	–0.91	0.3639	0.0103	0.30
SER	–0.55	–0.59	0.82	0.4852	–0.67	0.5010	0.0983	0.41
CAN	2.49	2.83	1.11	0.0402	2.24	0.0250	0.0000	2.08
CRN	0.06	0.08	1.30	0.9615	0.05	0.9632	0.0000	1.16
INF	0.19	0.17	0.60	0.7449	0.31	0.7544	0.0183	0.58
CPR	0.56	0.63	1.69	0.7274	0.33	0.7400	0.0000	0.89
PRE	1.08	1.26	0.88	0.2157	1.23	0.2185	0.0000	1.09

TYP	2.04	2.34	0.71	0.0138	2.87	0.0041	0.0000	1.41
FRA	0.72	0.77	1.13	0.5107	0.64	0.5236	0.0000	1.94
PO2	0.31	0.24	1.83	0.8468	0.17	0.8646	0.0000	1.51
PH	1.94	2.18	1.96	0.2948	0.99	0.3211	0.0000	1.46
PCO	–2.15	–2.21	1.58	0.1598	–1.36	0.1728	0.0000	1.64
BIC	–0.34	–0.14	2.05	0.8564	–0.16	0.8698	0.0000	3.36
CRE	0.49	0.60	1.82	0.7671	0.27	0.7882	0.0000	1.78
LOC	3.25	3.92	1.21	0.0339	2.67	0.0075	0.0000	10.71

9. General Remarks

When the data sets have large number of variables, the standard error computations are not reliable and hence the respective *p*-values. Bootstrap *p*-values should be preferred as they represent the true distribution of the estimates compared to the asymptotic distributions.

References

[1] D. R. Cox and E. J. Snell, Analysis of Binary Data, 2nd ed., Chapman and Hall, London, UK, 1989.

[2] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed., Wiley, New York, 2000.

[3] P. McCullagh and J. A. Nelder, Generalized Linear Models, 2nd ed., Chapman and Hall, New York, 1989.

[4] D. A. Powers and Y. Xie, Statistical Methods for Categorical Data Analysis, Academic Press, London, UK, 2000.

[5] F. L. Ramsey and D. W. Schafer, The Statistical Sleuth: A Course in Methods of Data Analysis, 2nd ed., Duxbury, Pacific Grove, CA, 2002.

[6] T. P. Ryan, Modern Regression Methods, Wiley, New York, 1997.