Journal of Undergraduate Research at Minnesota State University, Mankato

Volume 5 Article 22

2005

Bird Keeping and Lung Cancer

Andrew Tackmann
Minnesota State University, Mankato

Jonathan Hellman
Minnesota State University, Mankato

Jamie Johnson
Minnesota State University, Mankato

Bird Keeping and Lung Cancer

Andrew Tackmann andrew.tackmann@mnsu.edu
Jonathan Hellman jonathan.hellman@mnsu.edu
and
Jamie Johnson johnsj6@mnsu.edu

Faculty Mentor: Dr. Mezbahur Rahman mezbahur.rahman@mnsu.edu

Department of Mathematics and Statistics

Minnesota State University, Mankato, MN 56002, USA

Abstract

Logistic regression is reviewed in estimating parameters and in making inferences about the parameters. A contingency table approach in computing goodness of fit in logistic regression is elaborated. An existing data on a sample of lung cancer patients and a control group is used to apply the procedures discussed. The data reveals that between the groups considered, the factors ‘bird keeping’ and ‘the number of years of smoking’ are significant as the causes for lung cancer.

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 Logistic Regression, more specifically, logit regression. For the latest developments in the area, the reader is referred to Cox and Snell (1989), McCullagh and Nelder (1989), Ryan (1997), Hosmer and Lemeshow (2000), Powers and Xie (2000), and the references therein.

Let $X_1, X_2, \ldots, X_p$ be p different regressors and Y be a 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

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

where $0 \le \Lambda_i \le 1$ and $P(Y_i = 0) = 1 - \Lambda_i$. We define the logit regression model as

$$\Lambda(\mathbf{x}_i^T \boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_{pi} x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_{pi} x_{pi})} \tag{1}$$

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

In the following sections we describe the commonly used maximum likelihood estimation procedure, inferences about the parameters, the goodness of fit procedures, and the interpretations of the estimates through the odds ratios and the marginal effects. Then we apply the procedures for a data set. Finally, we give a brief conclusion about the findings in the data set and some general comments.

2 Estimation of Parameters

From (1), we have

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

where $\mathbf{x}^T = [1, x_1, x_2, x_3, \ldots, x_p]$ and 'ln' stands for the natural logarithm. For reasons described in Rahman et al. (2001), we wish to use the method of maximum likelihood to estimate the parameters of this model. 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

$$\mathrm{L} = \prod_i [\Lambda(\mathbf{x}_i^T\boldsymbol{\beta})]^{y_i}[1-\Lambda(\mathbf{x}_i^T\boldsymbol{\beta})]^{1-y_i},$$

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

$$\ln \mathrm{L} = \sum_i \{y_i \ln[\Lambda(\mathbf{x}_i^T\boldsymbol{\beta})] + (1-y_i)\ln[1-\Lambda(\mathbf{x}_i^T\boldsymbol{\beta})]\}. \tag{2}$$

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

$$\frac{\delta \ln(L(\boldsymbol{\beta}))}{\delta\boldsymbol{\beta}} = \mathbf{U}(\boldsymbol{\beta}) = \sum_i \{[y_i - \Lambda(\mathbf{x}_i^T\boldsymbol{\beta})]\mathbf{x}_i\}, \text{ and}$$

$$-\left[\frac{\delta^2 \ln(L(\boldsymbol{\beta}))}{\delta\boldsymbol{\beta}\boldsymbol{\beta}^T}\right] = \mathbf{I}(\boldsymbol{\beta}) = \sum_i \{\Lambda(\mathbf{x}_i^T\boldsymbol{\beta})[1-\Lambda(\mathbf{x}_i^T\boldsymbol{\beta})]\mathbf{x}_i\mathbf{x}_i^T\}. \tag{3}$$

At the k^{th} iteration, the estimates are obtained using the equation

$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k-1)} + \left[\mathbf{I}\left(\hat{\boldsymbol{\beta}}^{(k-1)}\right)\right]^{-1}\mathbf{U}\left(\hat{\boldsymbol{\beta}}^{(k-1)}\right),$$

DOI: 10.56816/2378-6949.1151

where $\hat{\boldsymbol{\beta}}^{(0)}$ is obtained by regressing the y on the x's. The iteration is stopped when the consecutive iteration values are close and/or the log-likelihood values are maximized (see Powers and Xie (2000) for details).

3 Inferences about the Parameters

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

$$Z_j = \frac{\hat{\beta}_j}{S\hat{E}(\hat{\beta}_j)}, \qquad j = 0,1,2,\ldots,p$$

where $S\hat{E}(\hat{\beta}_j) = \sqrt{\left(\left[\mathbf{I}(\hat{\boldsymbol{\beta}})\right]_{jj}^{-1}\right)}$, square root of the j$^{\text{th}}$ element of the inverse of $\mathbf{I}(\boldsymbol{\beta})$ in (3) evaluated at $\hat{\boldsymbol{\beta}}$. Then for a large sample Z_j will have approximate standard normal distribution under the null hypothesis, $H_0 : \beta_j = 0$. Significance for a set of parameters is discussed in section 4.1.

4 Goodness of Fit

Goodness of fit in logit model is different than usual linear regression models. The sum of squares and usual R^2 statistic do not have same interpretation as in linear regression. Here we discuss the commonly used likelihood ratio test for the reduced model compared with a presumed full model. Then we elaborate on a fairly new approach given by Hosmer and Lemeshow (1980), Chi-square statistic using a contingency table.

4.1 Likelihood ratio test

ln L in (2) can not be used alone as an index of fit because it is dependent on the size of the sample. Different values of ln L result when competing models differ in the number of parameters when fitted to the same data. The number of parameters, in general, should be more than one, and significantly less than the number of observations. To assess model fit, we need to know how one model fits relative to another. An indicator of model fit which measures the extent to which the current model deviates from a more generalized model is given by the likelihood-ratio statistic:

$$\chi^2_{\mathrm{R}} = -2\ln\left(\frac{L_c}{L_f}\right) = -2(\ln L_c - \ln L_f),$$

where $\ln L_c$ is the maximized log-likelihood for the current model, $\ln L_f$ is the maximized log-likelihood for the more generalized model, R stands for reduction in maximized log-likelihood. The likelihood ratio statistic χ^2_R has a Chi-Square distribution with $K_f - K_c$ degrees of freedom, where K_f and K_c denote the number of parameters in the full or the generalized model and the model under consideration, respectively. The test is upper-tailed. Small *p*-value indicates that the variables left out are significant. So, the higher the *p*-value the better the model is. A comparative study of different choices of general models can be seen in Simonoff (1998).

4.2 Contingency table approach

Hosmer and Lemeshow (1980) suggested a method of finding a goodness of fit statistic using the idea of χ^2 statistic in a contingency table. Here we will name it χ_C^2 as we have another Chi-Square statistic in Section 4.1. Using the estimated probabilities, the data can be divided into g groups. Hosmer and Lemeshow (2000, p.148) mentioned that the data be divided into 10 groups by dividing the possible estimated probabilities between 0 and 1 equally with 0.1 width of each interval which often violates usual assumption of expected frequencies in a cell of at least 5, Triola (2003, p. 273). We suggest that the number of groups should be such that the expected frequencies in each group is at least 5. Then there can be situations such that the degrees of freedom in χ_C^2 statistic is either zero or negative. In such situations, the χ_C^2 statistic should not be computed. The χ_C^2 statistic can be computed as

$$\chi_C^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}$$

where $E_i = \sum_{l=1}^{n_i} \Lambda_{il}$, n_i is the number of elements in the i^{th} group, and O_i is the observed number of successes or the number of one's in the respective group. The χ_C^2 statistic has approximate Chi-Square distribution with $g-(p+1)-1$ degrees of freedom. Hosmer and Lemeshow (2000) did not consider the fact that in estimating the expected frequencies, $p+1$ parameters are estimated and hence the degrees of freedom is reduced, see Madansky (1988). The smaller the χ_C^2 value is the better the fit is as the fitted values are closer to the observed values. Hence, if we calculate the *p*-value as $P\left(\chi_C^2 > \chi_C^2 \text{ observed}\right)$, higher the *p*-value is better the fit is.

5 Interpreting the Estimates

There are mainly two different ways one can interpret the estimates from the logit models. If the covariate is dichotomous or polychotomous, the interpretations could be made through finding the odds ratios. If the covariate is continuous or at least ordinal, the interpretations could be made by finding the marginal effects.

5.1 The odds-ratio

The odds of success is the ratio $\omega = \Lambda/(1-\Lambda)$, where Λ is the probability of success. If the covariate is dichotomous, we can code the values of the covariate as 1 or 0, then the odds ratio is the ratio of the odds for the two different values of the regression.
Say,

$$\theta = \frac{\omega_1}{\omega_0} = \frac{\Lambda_1/(1-\Lambda_1)}{\Lambda_0/(1-\Lambda_0)} = \exp(\beta_1),$$

where β_1 is unknown but fixed coefficient for the factor. If the covariate is polychotomous we replace the corresponding column by $s-1$ columns of 1's and 0's where s stands for the number of possible values for the covariate; '1' for the presence of that particular level and '0' for absence. Then the computation of odds-ratio for a specific level of the polychotomous factor is concerned is computed will be similar to dichotomous case as mentioned above. All inferences will also follow in similar way.

5.2 The marginal effects

The marginal effect expresses the change in the estimated prediction probability for a unit change in the independent variable. In the logit model, a unit change in x_i produces a β change in y_i. Thus, one could conceive of β as the marginal effect of x_i on y_i. The marginal effect of the $\mathrm{k}^{\underline{\mathrm{th}}}$ independent variable is given by

$$\theta_k = \frac{\delta \Pr(y_i = 1 \mid \mathbf{x}_i)}{\delta x_{ik}} = \frac{\delta \Lambda(\mathbf{x}_i^T \boldsymbol{\beta})}{\delta x_{ik}} = [\Lambda(\mathbf{x}_i^T \boldsymbol{\beta})][1-\Lambda(\mathbf{x}_i^T \boldsymbol{\beta})]\beta_k, \tag{4}$$

which is the rate of change in the success probability in the neighborhood of a particular value of x. Marginal effects associated with different values of x are useful. Substituting $\bar{\mathbf{x}}^T \boldsymbol{\beta}$ for $\mathbf{x}_i^T \boldsymbol{\beta}$ in equation (4) gives an average marginal effect associated with each independent variable, as opposed to each possible value of x. The marginal effects are parametric functions and hence inferences about them can be made through significance testing and by finding the confidence intervals. One can find estimates of the standard errors by using the method called the delta method as described in Ramsey and Schafer (2002, pp 328-329).

The estimate of θ_k is $\hat{\theta}_k$, the maximum likelihood estimate, where

$$\hat{\theta}_k = \hat{\Lambda}\left(\bar{\mathbf{x}}^T \hat{\boldsymbol{\beta}}\right)\left(1-\hat{\Lambda}\left(\bar{\mathbf{x}}^T \hat{\boldsymbol{\beta}}\right)\right)\hat{\beta}_k. \tag{5}$$

Let us consider the first derivatives of $\hat{\theta}_k$ with respect to β_i's evaluated at $\hat{\boldsymbol{\beta}}$. These are denoted as

$$\hat{\theta}'_k = \left.\frac{\delta\,\hat{\theta}_k}{\delta\beta_i}\right|_{\hat{\boldsymbol{\beta}}}, \qquad i = 0,1,2,\ldots,p.$$

Then the variance of $\hat{\theta}_k$ can be written as

$$\mathrm{Var}(\hat{\theta}_k) = \sum_{j=0}^{p}\sum_{l=0}^{p}\hat{\theta}'_{kj}\hat{\theta}'_{kl}\mathrm{Cov}(\hat{\beta}_j,\hat{\beta}_l)\,. \tag{6}$$

The estimates of the variances and covariances of the $\hat{\beta}$'s are the elements of $\left[\mathbf{I}(\hat{\boldsymbol{\beta}})\right]^{-1}$, where $\mathbf{I}(\hat{\boldsymbol{\beta}})$ is in (3). For the logit model,

$$\hat{\theta}'_{k0} = \frac{\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\left(1-\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\right)}{\left(1+\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\right)^3}\hat{\beta}_k, \tag{7}$$

$$\hat{\theta}'_{kk} = \frac{\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\left(1-\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\right)}{\left(1+\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\right)^3}\overline{x}_k\hat{\beta}_k + \frac{\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})}{\left(1+\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\right)^2},\ \ k \neq 0,\ \text{and} \tag{8}$$

$$\hat{\theta}'_{kk'} = \frac{\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\left(1-\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\right)}{\left(1+\exp(\overline{\mathbf{x}}^T\hat{\boldsymbol{\beta}})\right)^3}\overline{x}_{k'}\hat{\beta}_k, \tag{9}$$

where $k' \neq k, k \neq 0$ and are given by Heien et al. (2004).

6 Application

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, who were age 65 or younger, and who 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 (2002, Display 20.2). The data is displayed in the Appendix. 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

Let us rename the variables $LC = Y$, $FM = X_1$, $SS = X_2$, $BK = X_3$, $AG = X_4$, $YR = X_5$, and $CD = X_6$ to keep consistency with the notations in the theories described above.

There are six different factors in the data. We considered all $2^6 - 1 = 63$ possible models. The logit regression is considered as the response variable 'the presence of lung cancer' is binary. Maximum likelihood estimate of the parameters are computed as described in Section 2 and are given in Table 2. Then Z-statistics and the *p*-values are computed to test the significance for individual factors in the model as mentioned in Section 3 and are given in Table 3. We also computed the Chi-square statistics and their *p*-values as described in Section 4.1 and Section 4.2 and are given in Table 4.

By assuming that a considerable time and thoughts were given in selecting the factors in collecting data, we consider the model including all six factors as the full model. Among all 63 models, we wanted to concentrate on the ones having higher *p*-values for both the χ_C^2 and the χ_R^2 statistics. High *p*-value for χ_C^2 indicates that the observed frequencies and the expected frequencies are not significantly different. High *p*-value for χ_R^2 indicates that the variables left out are not significant. There are ten such models. We displayed necessary results for a total eleven models including the full model in Table 1 – Table 4. The models used are:

Table 1: Assumed Models

Model 1 $$\ln\left[\frac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Model 2 $$\ln\left[\frac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Model 3 $$\ln\left[\frac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Model 4 $$\ln\left[\frac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6$$

Model 5 $$\ln\left[\frac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Model 6 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6$

Model 7 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5$

Model 8 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5$

Model 9 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5$

Model 10 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$

Model 11 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = \beta_0 + \beta_3 x_3 + \beta_5 x_5 + \beta_6 x_6$

Table 2: Estimated Models

Model 1 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -1.9374 + 0.5613x_1 + 0.1054x_2 + 1.3626x_3 - 0.0398x_4 + 0.0729x_5 + 0.0260x_6$

Model 2 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -1.4815 + 0.0029x_2 + 1.4079x_3 - 0.0407x_4 + 0.0656x_5 + 0.0237x_6$

Model 3 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -1.9241 + 0.5366x_1 + 1.3521x_3 - 0.0388x_4 + 0.0719x_5 + 0.0263x_6$

Model 4 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -3.8094 + 0.5836x_1 + 0.0396x_2 + 1.4356x_3 + 0.0568x_5 + 0.0311x_6$

Model 5 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -1.4816 + 1.4075x_3 - 0.0407x_4 + 0.0656x_5 + 0.0238x_6$

Model 6 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -3.7806 + 0.5742x_1 + 1.4313x_3 + 0.0565x_5 + 0.0311x_6$

Model 7 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -3.5322 + 0.5304x_1 + 0.0622x_2 + 1.4208x_3 + 0.0659x_5$

Model 8 $\ln\left[\dfrac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -3.4978 + 0.5153x_1 + 0.0004x_3 + 0.0004x_5$

https://cornerstone.lib.mnsu.edu/jur/vol5/iss1/22
DOI: 10.56816/2378-6949.1151

Model 9 $$\ln\left[\frac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -3.1651 - 0.0368x_2 + 1.4702x_3 + 0.0582x_5$$

Model 10 $$\ln\left[\frac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -1.0336 + 1.3766x_3 - 0.0461x_4 + 0.0749x_5$$

Model 11 $$\ln\left[\frac{\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}{1-\Lambda(\mathbf{x}^T\hat{\boldsymbol{\beta}})}\right] = -3.4071 + 1.4959x_3 + 0.0493x_5 + 0.0284x_6$$

The model 1 is the full model since it has all of the factors in the data. The other ten models are chosen because of the high p-values for both χ_C^2 and χ_R^2 statistics. In Table 3, the p-values of the β's show that whether that parameter is significant in that particular model. Every time the factor 'bird keeping' was left in the model showed significant in that model. Similarly, the factor 'number of years of smoking', when left in the model was significant in that model. In Table 4, the p-values for χ_C^2 and χ_R^2 are given. Here we consider that high total p-value for χ_C^2 and χ_R^2 will lead to a better model. Model 3 has the highest total p-value of 0.8778. But χ_C^2 has small p-value of 0.0558. So, we consider model 6 is the best model with total p-value of 0.8538 and none of the two p-values are small. It is to be noted that in none of the 63 models, the two variables 'social status' and 'age' are significant.

Table 3: Parameter Significance

Models	Parameter Estimators	Estimated Standard Error	p-values	Models	Parameter Estimators	Estimated Standard Error	p-values
Model 1	$\hat{\beta}_0$	1.8043	0.2829		$\hat{\beta}_5$	0.0248	0.0082
	$\hat{\beta}_1$	0.5312	0.2907		$\hat{\beta}_6$	0.0250	0.3430
	$\hat{\beta}_2$	0.4688	0.8221	Model 6	$\hat{\beta}_0$	0.7847	0.0000
	$\hat{\beta}_3$	0.4113	0.0009		$\hat{\beta}_1$	0.5155	0.2653
	$\hat{\beta}_4$	0.0355	0.2625		$\hat{\beta}_3$	0.4033	0.0004
	$\hat{\beta}_5$	0.0265	0.0059		$\hat{\beta}_5$	0.0201	0.0050
	$\hat{\beta}_6$	0.0255	0.3081		$\hat{\beta}_6$	0.0247	0.2080
Model 2	$\hat{\beta}_0$	1.7295	0.3917	Model 7	$\hat{\beta}_0$	0.7716	0.0000
	$\hat{\beta}_2$	0.4570	0.9950		$\hat{\beta}_1$	0.5236	0.3111
	$\hat{\beta}_3$	0.4077	0.0006		$\hat{\beta}_2$	0.4585	0.8921
	$\hat{\beta}_4$	0.0349	0.2438		$\hat{\beta}_3$	0.4047	0.0004
	$\hat{\beta}_5$	0.0250	0.0087		$\hat{\beta}_5$	0.0188	0.0005

	$\hat{\beta}_6$	0.0251	0.3446	Model 8	$\hat{\beta}_0$	0.7282	0.0000
Model 3	$\hat{\beta}_0$	1.8067	0.2869		$\hat{\beta}_1$	0.5116	0.3139
	$\hat{\beta}_1$	0.5196	0.3018		$\hat{\beta}_3$	0.4011	0.0004
	$\hat{\beta}_3$	0.4084	0.0009		$\hat{\beta}_5$	0.0187	0.0004
	$\hat{\beta}_4$	0.0353	0.2709	Model 9	$\hat{\beta}_0$	0.6622	0.0000
	$\hat{\beta}_5$	0.0261	0.0059		$\hat{\beta}_2$	0.4466	0.9343
	$\hat{\beta}_6$	0.0255	0.3011		$\hat{\beta}_3$	0.4009	0.0002
Model 4	$\hat{\beta}_0$	0.8224	0.0000		$\hat{\beta}_5$	0.0169	0.0006
	$\hat{\beta}_1$	0.5273	0.2683	Model 10	$\hat{\beta}_0$	1.6607	0.5337
	$\hat{\beta}_2$	0.4645	0.9321		$\hat{\beta}_3$	0.4007	0.0006
	$\hat{\beta}_3$	0.4066	0.0004		$\hat{\beta}_4$	0.0343	0.1790
	$\hat{\beta}_5$	0.0203	0.0052		$\hat{\beta}_5$	0.0230	0.0011
	$\hat{\beta}_6$	0.0248	0.2096	Model 11	$\hat{\beta}_0$	0.6868	0.0000
Model 5	$\hat{\beta}_0$	1.7295	0.3916		$\hat{\beta}_3$	0.3982	0.0002
	$\hat{\beta}_3$	0.4037	0.0005		$\hat{\beta}_5$	0.0187	0.0084
	$\hat{\beta}_4$	0.0347	0.2407		$\hat{\beta}_6$	0.0244	0.2435

Table 4: Goodness of Fit

Model	χ_c^2	*df*	*p*-value	χ_R^2	*df*	*p*-value
1	4.4765	1	0.0343	-	-	-
2	2.9506	1	0.0858	1.1184	1	0.2903
3	3.6590	1	0.0558	0.0505	1	0.8220
4	0.7963	1	0.3722	1.2961	1	0.2549
5	3.5497	1	0.0596	1.1184	2	0.5717
6	0.9386	1	0.3326	1.3033	2	0.5212
7	1.7434	1	0.1867	2.8806	2	0.2369
8	2.8924	1	0.0890	2.8990	3	0.4075
9	2.9601	1	0.0853	3.9092	3	0.2714
10	2.7769	1	0.0956	2.0185	3	0.5686
11	1.3691	1	0.2420	2.5490	3	0.4665

In interpreting the estimates in model 6, using the methods described in Section 5, we need the estimates of the variance covariance matrix and displayed in Table 5.

Table 5: Variance-Covariance Matrix for Model 6

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_3$	$\hat{\beta}_5$	$\hat{\beta}_6$
$\hat{\beta}_0$	0.6158	-0.1892	-0.1151	-0.0112	-0.0065
$\hat{\beta}_1$	-0.1892	0.2657	-0.0242	0.0036	0.0014
$\hat{\beta}_3$	-0.1151	-0.0242	0.1627	0.0006	0.0007
$\hat{\beta}_5$	-0.0112	0.0036	0.0006	0.0004	-0.0002
$\hat{\beta}_6$	-0.0065	0.0014	0.0007	-0.0002	0.0006

The estimated odds ratio for the variable FM (X_1) is $\exp(\hat{\beta}_1) = \exp(0.5742) = 1.78$. That is females are 1.78 times more likely to have lung cancer than their counter parts. The 95% confidence interval for the odds ratio for FM (X_1) is computed as

$$\exp(\hat{\beta}_1 \pm z_{\alpha/2} S\hat{E}(\hat{\beta}_1)) = \exp(0.5742 \pm 1.96\sqrt{0.2657}) = (0.6466, 4.8769).$$

The estimated odds ratio for the variable BK (X_3) is $\exp(\hat{\beta}_3) = \exp(1.4313) = 4.18$. That is, the person who keeps bird in this target population is 4.18 times more likely to have lung cancer than who do not. The 95% confidence interval for the odds ratio for FM (X_3) is computed as

$$\exp(\hat{\beta}_3 \pm z_{\alpha/2} S\hat{E}(\hat{\beta}_3)) = \exp(1.4313 \pm 1.96\sqrt{0.1627}) = (1.8978, 9.2248).$$

The marginal effects for the continuous covariates YR (X_5) and CD (X_6) are computed using equation (4), as follows:

$\hat{\Lambda} = \dfrac{\exp(-3.7806 + 0.5742\bar{x}_1 + 1.4313\bar{x}_3 + 0.0565\bar{x}_5 + 0.0311\bar{x}_6)}{1 + \exp(-3.7806 + 0.5742\bar{x}_1 + 1.4313\bar{x}_3 + 0.0565\bar{x}_5 + 0.0311\bar{x}_6)}$, where $\bar{x}_1$=0.2449, $\bar{x}_3$=0.4558, $\bar{x}_5$=27.8503, and $\bar{x}_6$=15.7483.

$\hat{\theta}_i = \hat{\Lambda}(1-\hat{\Lambda})\hat{\beta}_i$, $\quad i = 5,6. \quad \hat{\theta}_5 = 0.0115, \quad \hat{\theta}_6 = 0.0063$.

$\hat{\theta}_5$ = 0.0115 indicates that if the number of years of smoking is increased by one year, then the probability of having lung cancer is increased by 0.0115. Similarly, $\hat{\theta}_6$ = 0.0063 indicates that the probability of lung cancer increases by 0.0063 for a unit increase in the number of cigarettes per day. The 95% confidence intervals for θ_5 and θ_6 are computed as $\hat{\theta}_i \pm z_{\alpha/2} S\hat{E}(\hat{\theta}_i)$ where $S\hat{E}(\hat{\theta}_i) = \sqrt{V\hat{a}r(\hat{\theta}_k)}$ is equation (6). In Table 6, we give the estimates of the partial derivatives in (7), (8), and (9) needed in computing $S\hat{E}(\hat{\theta}_i)$.

Table 6: Partial Derivatives for Model 6

$\hat{\theta}'_{50}$	0.0054	$\hat{\theta}'_{60}$	0.0030

$\hat{\theta}'_{51}$	0.0031	$\hat{\theta}'_{61}$	0.0017
$\hat{\theta}'_{53}$	0.0078	$\hat{\theta}'_{63}$	0.0043
$\hat{\theta}'_{55}$	0.0003	$\hat{\theta}'_{65}$	0.0003
$\hat{\theta}'_{56}$	0.0002	$\hat{\theta}'_{66}$	0.0001

Using (6), we get, $S\hat{E}(\hat{\theta}_5)) = 0.0035$ and $S\hat{E}(\hat{\theta}_6)) = 0.0020$. Then the 95% confidence intervals for θ_5 and θ_6 are respectively, (0.0046, 0.0184) and (0.0024, 0.0102).

Interpretations for the estimates in the full model might also be of interest and are given below. It is to be noted that the variables, gender (FM), social status (SS), and bird keeping (BK) are categorical. Age (AG), number of years of smoking (YR), and number of cigarettes per day (CD) are continuous measurements. So, the odds ratios are computed for the variables gender, social status, and bird keeping are
$\exp(\hat{\beta}_1) = \exp(0.5613) = 1.75$, $\exp(\hat{\beta}_2) = \exp(0.1054) = 1.11$, and
$\exp(\hat{\beta}_3) = \exp(1.3626) = 3.91$. Females are 1.75 times more vulnerable for lung cancer than male, higher class population is 1.11 times more vulnerable for lung cancer than the lower class, and bird keepers are 3.91 times more vulnerable for lung cancer than who are not. Corresponding 95% confidence intervals are
$\exp(\hat{\beta}_1 \pm z_{\alpha/2} S\hat{E}(\hat{\beta}_1)) = \exp(0.5613 \pm 1.96\sqrt{0.2821}) = (0.6190, 4.9645)$,
$\exp(\hat{\beta}_2 \pm z_{\alpha/2} S\hat{E}(\hat{\beta}_2)) = \exp(0.1054 \pm 1.96\sqrt{0.2198}) = (0.4433, 2.7852)$, and
$\exp(\hat{\beta}_3 \pm z_{\alpha/2} S\hat{E}(\hat{\beta}_3)) = \exp(1.3626 \pm 1.96\sqrt{0.1627}) = (1.7448, 8.7459)$.

Table 7: Variance-Covariance Matrix for the Full Model

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
$\hat{\beta}_0$	3.2553	-0.2344	-0.0303	-0.2172	-0.0564	0.0124	-0.0134
$\hat{\beta}_1$	-0.2344	0.2821	0.0520	-0.0171	0.0003	0.0040	0.0012
$\hat{\beta}_2$	-0.0303	0.0520	0.2198	0.0233	-0.0019	0.0021	-0.0006
$\hat{\beta}_3$	-0.2172	-0.0171	0.0233	0.1691	0.0019	-0.0002	0.0010
$\hat{\beta}_4$	-0.0564	0.0003	-0.0019	0.0019	0.0013	-0.0006	0.0002
$\hat{\beta}_5$	0.0124	0.0040	0.0021	-0.0002	-0.0006	0.0007	-0.0002
$\hat{\beta}_6$	-0.0134	0.0012	-0.0006	0.0010	0.0002	-0.0002	0.0007

The marginal effects for the variables age, number of years of smoking, and number of cigarettes per day are computed below.

$$\hat{\Lambda} = \frac{\exp(-1.9374 + 0.5613\bar{x}_1 + 0.1054\bar{x}_2 + 1.3626\bar{x}_3 - 0.0398\bar{x}_4 + 0.0729\bar{x}_5 + 0.0260\bar{x}_6)}{1 + \exp(-1.9374 + 0.5613\bar{x}_1 + 0.1054\bar{x}_2 + 1.3626\bar{x}_3 - 0.0398\bar{x}_4 + 0.0729\bar{x}_5 + 0.0260\bar{x}_6)},$$

where $\bar{x}_1$=0.2449, $\bar{x}_2$=0.3061, $\bar{x}_3$=0.4558, $\bar{x}_4$=56.9660, $\bar{x}_5$=27.8503, and $\bar{x}_6$=15.7483.

$\hat{\theta}_i = \hat{\Lambda}(1-\hat{\Lambda})\hat{\beta}_i$, $i = 4,5,6$. $\hat{\theta}_4 = -0.0079$, $\hat{\theta}_5 = 0.0145$, $\hat{\theta}_6 = 0.0052$. $\hat{\theta}_4 = -0.0079$ indicates that age is increased by one year the probability of having lung cancer is reduced by 0.0079. Since the data is collected for older people only, people who survived without having lung cancer have lower risk of having lung cancer in the near future. $\hat{\theta}_5 =$ 0.0145 indicates that if the number of years of smoking is increased by one year, then the probability of having lung cancer is increased by 0.0145. Similarly, $\hat{\theta}_6 = 0.0052$ indicates that the probability of lung cancer increases by 0.0052 for a unit increase in the number of cigarettes per day. The 95% confidence intervals for θ_4, θ_5 and θ_6 are computed as $\hat{\theta}_i \pm z_{\alpha/2}\hat{SE}(\hat{\theta}_i)$ where $\hat{SE}(\hat{\theta}_i) = \sqrt{V\hat{a}r(\hat{\theta}_k)}$ is equation (6). In Table 8, we give the estimates of the partial derivatives in equations (7), (8), and (9) needed in computing $\hat{SE}(\hat{\theta}_i)$.

Table 8: Partial Derivatives for the Full Model

$\hat{\theta}'_{40}$	-0.0036	$\hat{\theta}'_{50}$	0.0066	$\hat{\theta}'_{60}$	0.0023
$\hat{\theta}'_{41}$	-0.0009	$\hat{\theta}'_{51}$	0.0016	$\hat{\theta}'_{61}$	0.0006
$\hat{\theta}'_{42}$	-0.0011	$\hat{\theta}'_{52}$	0.0020	$\hat{\theta}'_{62}$	0.0007
$\hat{\theta}'_{43}$	-0.0016	$\hat{\theta}'_{53}$	0.0030	$\hat{\theta}'_{63}$	0.0011
$\hat{\theta}'_{44}$	-0.0049	$\hat{\theta}'_{54}$	0.3734	$\hat{\theta}'_{64}$	0.1332
$\hat{\theta}'_{45}$	-0.0997	$\hat{\theta}'_{55}$	0.3815	$\hat{\theta}'_{65}$	0.0651
$\hat{\theta}'_{46}$	-0.0564	$\hat{\theta}'_{56}$	0.1032	$\hat{\theta}'_{66}$	0.2357

Using equation (6), we get, $\hat{SE}(\hat{\theta}_4)) = 0.0069$, $\hat{SE}(\hat{\theta}_5)) = 0.0048$ and $\hat{SE}(\hat{\theta}_6)) = 0.0051$. Then the 95% confidence intervals for θ_4, θ_5 and θ_6 are respectively (-0.0214,0.0056), (0.0051, 0.0239) and (-0.0048, 0.0152).

7. General Comments

The procedures mentioned in this paper can be applied to a wide range of categorical response models. Some software has automated logistic regression computations. Here we used MATLAB mathematical computational software to program for all the computations. Interested persons are encouraged to contact the first author to get hold of the MATLAB code. Depending on the number of classes considered in computing the χ^2_C statistic, the *p*-value varies significantly. In some single variate models, we were not able to compute *p*-values as the degrees of freedom became zero or negative.

The high odds ratio value for female versus male (1.78) reinforces the recent observations in biomedical research that gender should be taken into consideration in developing disease preventions.

Very high odds ratio value for the factor bird keeping (4.18) is alarming but should not be taken as its face value. We want to mention some drawbacks of the data. This data is from only one city and hence not to be generalized. The investigators might also have overlooked some key factors for lung cancer in the city from which the data was collected.

8. References

Cox, D. R. and Snell, E. J. (1989), *Analysis of Binary Data*, 2nd edition. London: Chapman and Hall.

Heien, Herbert C., Baumann, William A. and Rahman, Mezbahur (2004), An Application of Logistic Regression in Estimation and in Prediction, *Pakistan Journal of Statistics,* Vol. 20 (1), pp. 37-47.

Hosmer, David W. and Lemeshow, Stanley (1980), A Goodness-of-Fit Test for the Multiple Logistic Regression Model, Communications in Statistics, Vol. A10, 1043-1069.

Hosmer, David W. and Lemeshow, Stanley (2000), Applied Logistic Regression, 2nd edition. New York: John Wiley and Sons, Inc.

Madansky, Albert (1988), *Prescriptions for Working Statisticians.* Springer-Verlag.

McCullagh, P. and J. A. Nelder (1989), *Generalized Linear Models,* 2nd edition. New York: Chapman and Hall.

Powers, Daniel A. and Xie, Yu (2000), *Statistical Methods for Categorical Data Analysis*. San Diego, CA: Academic Press.

Rahman, Mezbuhar, Cortes, Judy and Pardis, Cyrus (2001), A Note on Logistic Regression, *Journal of Statistics & Management Systems*, Vol. 4, No. 2, pp. 175-187.

Ramsey, Fred L. and Schafer, Daniel W. (2002), *The Statistical Sleuth: A Course in Methods of Data Analysis*, 2nd edition. Pacific Grove, CA: Duxbury.

Ryan, Thomas P. (1997), *Modern Regression Methods*, New York: Wiley.

Siminoff, Jeffrey S. (1998), Logistic Regression, Categorical Predictors, and Goodness-of-Fit: It Depends on Who You Ask, *The American Statistician*, Vol. 52, No.1, pp. 10-14.

Triola, Mario F. (2004), 9th edition, Elementary Statistics. New York: Pearson Education, Inc.

Appendix

Data Table

SN	LC	FM	SS	BK	AG	YR	CD	SN	LC	FM	SS	BK	AG	YR	CD
1	1	0	0	1	37	19	12	75	0	0	0	0	57	35	15
2	1	0	0	1	41	22	15	76	0	0	0	0	57	24	15
3	1	0	1	0	43	19	15	77	0	0	0	0	58	38	20
4	1	0	0	1	46	24	15	78	0	0	0	0	58	39	20
5	1	0	0	1	49	31	20	79	0	0	0	0	58	22	10
6	1	0	1	0	51	24	15	80	0	0	0	0	58	15	40
7	1	0	1	1	52	31	20	81	0	0	0	1	59	36	15
8	1	0	1	0	53	33	20	82	0	0	0	1	59	35	20
9	1	0	0	1	56	33	10	83	0	0	1	0	59	41	12
10	1	0	0	0	56	26	25	84	0	0	0	0	59	37	15
11	1	0	0	0	56	35	40	85	0	0	1	0	59	7	1
12	1	0	0	0	56	36	25	86	0	0	0	0	59	34	20
13	1	0	0	1	56	36	20	87	0	0	1	1	60	25	15
14	1	0	0	1	57	39	25	88	0	0	1	1	60	39	12
15	1	0	0	0	58	38	20	89	0	0	1	0	60	34	1
16	1	0	0	0	58	35	25	90	0	0	1	0	60	0	0
17	1	0	1	0	58	42	30	91	0	0	0	0	61	42	12
18	1	0	0	1	59	39	20	92	0	0	0	0	61	43	20
19	1	0	1	1	59	40	15	93	0	0	0	1	62	38	20
20	1	0	0	1	60	38	15	94	0	0	1	0	62	0	0
21	1	0	0	0	61	28	15	95	0	0	1	0	62	14	30
22	1	0	1	1	62	39	20	96	0	0	0	0	62	44	30
23	1	0	0	0	62	43	20	97	0	0	0	0	62	28	18
24	1	0	0	1	62	40	15	98	0	0	1	1	63	0	0
25	1	0	1	1	63	41	40	99	0	0	0	1	63	0	0
26	1	0	0	0	63	45	20	100	0	0	0	0	63	22	15
27	1	0	0	1	63	41	10	101	0	0	0	0	63	22	20
28	1	0	0	1	64	42	20	102	0	0	1	0	63	41	20
29	1	0	0	1	64	44	15	103	0	0	0	0	63	20	15
30	1	0	1	1	64	47	16	104	0	0	1	0	63	43	20
31	1	0	1	1	64	13	30	105	0	0	1	0	63	42	10
32	1	0	0	1	64	42	20	106	0	0	0	1	64	40	20
33	1	0	0	0	64	32	3	107	0	0	0	1	64	40	10
34	1	0	1	0	65	45	10	108	0	0	0	1	64	46	20
35	1	0	1	1	65	43	30	109	0	0	0	0	64	41	6
36	1	0	0	1	66	50	25	110	0	0	1	0	64	39	25
37	1	0	0	1	66	47	10	111	0	0	1	0	64	39	20
38	1	1	0	1	44	22	15	112	0	0	0	0	64	45	20
39	1	1	0	1	46	24	15	113	0	0	0	0	64	36	15
40	1	1	0	1	47	25	25	114	0	0	0	0	64	44	20
41	1	1	0	1	49	27	20	115	0	0	0	0	65	44	6
42	1	1	0	1	49	23	20	116	0	0	1	0	65	30	20
43	1	1	0	1	50	28	20	117	0	0	1	0	65	47	45

44	1	1	0	1	54	33	6	118	0	0	0	0	65	46	20
45	1	1	0	1	58	37	20	119	0	0	0	0	65	34	10
46	1	1	0	1	61	38	15	120	0	0	0	1	66	38	25
47	1	1	0	1	62	0	0	121	0	0	1	0	66	42	18
48	1	1	0	0	63	29	20	122	0	0	0	0	66	0	0
49	1	1	0	0	64	40	25	123	0	0	1	0	67	0	0
50	0	0	0	0	37	16	2	124	0	1	0	0	43	21	20
51	0	0	0	1	38	20	20	125	0	1	0	1	45	0	0
52	0	0	1	0	40	13	25	126	0	1	0	1	46	24	4
53	0	0	1	1	42	21	8	127	0	1	0	1	46	0	0
54	0	0	1	0	42	17	15	128	0	1	1	1	46	16	5
55	0	0	0	0	43	25	25	129	0	1	1	0	47	0	0
56	0	0	1	0	46	24	20	130	0	1	0	1	49	25	15
57	0	0	0	1	47	28	40	131	0	1	0	1	49	0	0
58	0	0	0	1	49	28	10	132	0	1	0	0	49	25	15
59	0	0	0	1	49	15	10	133	0	1	0	0	49	27	20
60	0	0	1	0	51	5	4	134	0	1	0	1	50	0	0
61	0	0	1	0	51	17	10	135	0	1	0	0	51	23	12
62	0	0	1	1	52	30	37	136	0	1	0	1	53	9	10
63	0	0	0	1	52	28	25	137	0	1	1	0	55	24	15
64	0	0	1	0	53	29	10	138	0	1	0	1	58	34	15
65	0	0	0	0	53	19	15	139	0	1	0	1	60	36	25
66	0	0	0	1	55	39	15	140	0	1	0	0	60	0	0
67	0	0	0	0	55	41	30	141	0	1	0	0	61	0	0
68	0	0	0	0	55	18	10	142	0	1	1	1	62	0	0
69	0	0	0	1	56	36	20	143	0	1	0	1	62	20	10
70	0	0	0	1	56	22	25	144	0	1	1	0	63	0	0
71	0	0	0	0	56	39	25	145	0	1	0	0	64	39	20
72	0	0	0	0	56	32	30	146	0	1	1	0	64	0	0
73	0	0	1	0	56	19	5	147	0	1	0	0	65	7	2
74	0	0	0	1	57	24	8								

Authors' Biographies: ??

Faculty Mentor's Biography: Dr. Mezbahur Rahman, Associate Professor, Department of Mathematics and Statistics, Minnesota State University has obtained Ph.D. in Applied Statistics from the University of California Riverside in 1995. Dr. Rahman's research interest is in the areas of theoretical and applied statistics. He has twenty five different papers published in different journals to the date. He has refereed for twelve different journals.