*Adv. Appl. Res., Vol.3, No.2, (2011) pp 163 - 170*

# Logistic Regression for Classification in Agricultural Ergonomics

**Arpan Bhowmik[1], V. Ramasubramanian[1*], Chandrahas[2] and Adarsh Kumar[3]**

[1]Division of Biometrics and Statistical Modelling, [2]Division of Forecasting Techniques, [3]Division of Agricultural Engineering, Indian Agricultural Statistics Research Institute, Pusa, New Delhi – 110 012, India

[*]*Corresponding author: E Mail: ram_stat@yahoo.co.in*



## Abstract

Classification and prediction in agricultural systems are quite useful for effective planning. In this paper, logistic regression modeling has been employed for classification purposes on data pertaining to the area of agricultural ergonomics. Presence or absence of discomfort for the farm labourers in operating farm machineries has been considered as the dependent variable and associated quantitative and qualitative variables as regressors. From the different possible subsets of regressors, appropriate logistic regression models that best describe the dependent variable have been selected. Appropriate goodness of fit and predictive ability measures have been utilized for evaluating the performance of the fitted models. A single best regressor i.e., load given to the farm machinery during operation has been identified by employing variable selection based on collinearity diagnostics and stepwise logistic regression. Results of classifications of the test datasets revealed that logistic regression performed better than the conventionally used discriminant function analysis approach. The study revealed that logistic regression modeling can be employed as a viable alternative for classification purposes in the field of agricultural ergonomics (**Keywords:** Classificatory power, Hosmer and Lameshow goodness of fit, predictive ability, discriminant function)

## Introduction

Classification and prediction in agricultural systems are quite useful for effective planning. For this purpose various statistical approaches are in vogue. In this study, an attempt has been made to study logistic regression modeling with the aim of utilizing the same for classification purposes in the field of agricultural ergonomics. For

analysis of binary (dichotomous) responses such as presence or absence of discomfort for farm labourers during farm operations, logistic regression modeling which can be reformulated as a classification technique by considering the two distinct responses as two groups in the lines of discriminant function analysis wherein distinct observations with pre-defined group memberships along with associated variables are analysed and separated and new objects are allocated to the previously defined groups. Logistic regression as a classification tool has been widely used in various fields such as economics, medical science (epidemiology and health), psychology, classical ergonomics etc. To cite a few relevant references, Johnson *et al.* (1996) described the relationships between weather and outbreaks of potato late blight in the semi arid environment of south-central Washington with linear discriminant and logistic regression analyses and forecasted late blight outbreaks. Vergara and Page (2002) classified lumbar discomfort/absence of discomfort by relating with back posture and mobility in sitting-posture using both discriminant analysis and logistic regression. Gent *et al.* (2003) used logistic regression for classifying the geographical regions of origin of *Xanthomonas* strains. Mila *et al.* (2004) used logistic regression to estimate the probability of soybean Sclerotinia stem rot prevalence in north-central region of the United States using tillage practice, soil texture and weather variables (monthly air temperature and monthly precipitation from April to August) as inputs.

So far not much work has been done on application of logistic regression in the field of agricultural ergonomics in India. Large number of farm machines and hand tools used in Indian agriculture require involvement of human energy for operating in different modes (pedal, bicycle, flywheel etc.) for low energy power generation. The working environment in operating the different machines is strenuous for farm labourers and to quantify the discomfort involved modelling from statistical point of view is necessary. The commonly used body parts for manual power are upper limbs and legs depending upon the machine and are performed for its operation at a greater physiological cost and postural stress leading to discomfort and fatigue depending upon posture, force application, quantum and frequency. The farm labourers experience discomfort in hands and legs in general and thighs, knees, feet, legs, back, palms, buttock etc. in particular. Moreover, the human efficiencies vary at different loadings (given for differential mechanical output conditions) and body weight supported with the modes of operation. In this paper logistic regression and

discriminant analysis approaches have been applied to classify discomfort level of farm labourers so that effective measures can be made to rectify the .discomfort causing features of the farm machinery. For estimating the parameters involved in various logistic regression models, Maximum Likelihood Estimation (MLE) method was used. The fitted models have been used for classifying new observations not included in model fitting. Relevant goodness of fit measures has been utilized for assessing the adequacy of the fitted models.

## Materials and Methods

Indian farm employs 225 million workers, constituting 10 per cent of total world's workforce in agricultural activities (Ram *et al.,* 2008). Working environment of farm is labour intensive and strenuous. The data for the present study has been taken in the area of Agricultural Ergonomics from the Division of Agricultural Engineering, I.A.R.I., New Delhi collected during 2007 - 2008. The dependent variable considered for the present study was overall discomfort of the farm labourers during farm operation (Y) having only two levels 0 and 1 depending upon whether discomfort is absent or present. The set of explanatory variables considered are:

Quantitative: $X_1$ - load given to farm machineries during operation

$X_2$ - difference between working and resting heart rates

$X_3$ - oxygen consumption at the time of farm operation

Qualitative: $X_4$ - Mode of operation

$X_5$ - Body part discomfort

and $X_6$ – Percentage (%) of aerobic capacity of the farm labourers during operation

For the variable $X_1$ i.e., load given to the farm machinery five levels were there *viz*., no load, 0.90W, 1.80W, 2.70W and 3.60W (W is the unit of power i.e., Watt). For the variable $X_4$, two mutually exclusive levels *viz*., predominantly foot operated (e.g., Bicycle, Stepper and Pedal etc.) and other mode of operation (e.g., Flywheel, Rocking etc.) were considered. Thus, $X_4$ was a nominal variable. The ordinal variable $X_5$ i.e., Body Part Discomfort (BPD) has been considered at three levels as low, medium and high and was represented by two indicator variables ($z_1$, $z_2$) taking values (0, 0), (1, 0) and (0, 1) respectively. Passmore and Durrnin (1955) and Saha *et al*.

(1979) suggested a limit of 30 % to 35 % aerobic capacity of farm labourers for continuous 8 h work as an acceptable workload. Accordingly, levels of $X_6$ have been taken as low and high denoted '0' and '1' respectively *viz*., below 35 % and over 35 % of aerobic capacity of the farm labourers. The response in relation to the set of explanatory variables (six in number) has been observed upon nine subjects (farm labourers) with observations taken at three independent time periods under each of the five loads of operation considered. Thus, in total, $9 \times 3 \times 5 = 135$ observations were available for the study.

Out of the 135 observations available, 80 % i.e., 108 observations have been selected randomly for model fitting. The remaining 20 % i.e., 27 observations have been considered as test data set for model validation. Considering the set of regressors (both quantitative and qualitative), based on all possible subsets of variables, appropriate datasets and hence models have also been identified by employing variable selection based on collinearity diagnostics (considering quantitative variables only) and also by using stepwise logistic regression (considering both types of variables). For collinearity diagnostics, the detection measures considered are pairwise correlation coefficients among the variables, variance inflation factor (VIF) of each variable and Belsley's procedure of finding proportions of variance accounted for by each variable. After variable selection, various logistic regression models were developed separately based on 108 datapoints using MLE procedure.

Let Y denotes the binary response variable and $X_1, X_2, \ldots, X_r$ be the set of r explanatory variables. If $\pi$ is the probability that the event occurs, then $\pi = P(Y=1)$, assuming Y to take on value '1' when such an event occurs. Thus the simple binary logistic regression model is given by:

$$\begin{aligned} \pi &= P(Y=1|X_1 = x_1 \ldots X_r = x_r) \\ &= 1/(1+e^{-z}) \end{aligned} \qquad (1)$$

Where $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r$ and $\beta_0, \beta_1 \ldots \beta_r$ are the model parameters. Based on this binary logistic regression model, various logistic regression models have been fitted. For each of these fitted models, P(Y=1) has been calculated. These probabilities were used for classifying the observations in test data sets wherein if $P(Y=1) > 0.5$, then a value '1' has been assigned to Y to indicate the presence of discomfort of the farm laborer during farm operation otherwise, the value '0'. Once

models are fitted, testing of goodness of fit of logistic regression models has been done by using Hosmer-Lemeshow goodness-of-fit statistic which is the most common tool used in logistic regression analysis. Here, to start with**,** the observations are sorted in increasing order of their estimated event probability. The observations are then divided into approximately ten groups on the basis of the estimated probabilities. The number of groups may be smaller than 10 but there must be at least three groups in order that the Hosmer-Lemeshow statistic can be computed. Beside goodness-of-fit of logistic regression models, based on 108 datapoints the predictive ability of these logistic regression models have also been judged through Gamma and Somers' D statistics with higher values of these two indicating higher predictive ability of models. Performance of various logistic regression models (involving only continuous explanatory variables) has been compared with the corresponding results obtained from discriminant function analysis method through (2 × 2) classification tables. For comparison purposes the following measures have been used:

***Hit rate:*** Number of correct predictions divided by sample size.

***Sensitivity:*** Percent of correct predictions in the reference category (usually 1) of the dependent variable.

***Specificity:*** Percent of correct predictions in the given category (usually 0) of the dependent variable.

***False positive rate:*** It is the proportion of predicted event responses that were observed as non-events

***False negative rate:*** It is the proportion of predicted non-event responses that were observed as events.

## Results and Discussion

For collinearity diagnostics, pairwise correlations, variance inflation factors (VIF) for each variable and Belsley's variance decomposition proportions has been computed for the available quantitative explanatory variables and results are summarized in Table 1. From this table, it can be seen that VIF is maximum for $X_1$ and the conditional index for the three quantitative variables considered for the present study i.e., $X_1$, $X_2$ and $X_3$ are 1.00, 3.50 and 4.05 respectively. So, corresponding to the maximum conditional index i.e., 4.05, the variance decomposition proportion is highest for $X_1$. Correlation between $X_1$ and $X_2$ is 0.833 and the same between $X_1$ and

$X_3$ is 0.818. Both these correlations are higher as compared to the correlation between $X_2$ and $X_3$ which is equal to 0.787. Thus variable $X_1$ is the most influential variable, which is nothing but load given to farm machineries during operation.

**Table 1.** Collinearity diagnostics (Intercept adjusted model)

| Variables | Variance Inflation Factor (VIF) | Eigen values | Conditional index | Proportion of variances | | |
|---|---|---|---|---|---|---|
| | | | | $X_1$ | $X_2$ | $X_3$ |
| $X_1$ | 4.22 | 2.63 | 1.00 | 0.03 | 0.04 | 0.04 |
| $X_2$ | 3.67 | 0.22 | 3.50 | 0.02 | 0.50 | 0.81 |
| $X_3$ | 3.89 | 0.16 | 4.05 | 0.95 | 0.47 | 0.16 |



Beside collinearity diagnostics, stepwise logistic regression procedure for both quantitative and qualitative variables was performed in which variable selection and model building are done simultaneously. By stepwise logistic regression, the final model consisted of only $X_1$ apart from intercept. It is to be noted here that, in both the procedures employed, the selected variable is found to be $X_1$. The results obtained by considering all possible subsets have been discussed subsequently which also include the model corresponding to variable $X_1$, hence the results of the fitted model by considering both collinearity diagnostics and stepwise logistic regression has not been presented separately here. As there are six explanatory variables $X_1$, $X_2$, $X_3$, $X_4$, $X_5$ and $X_6$ in the present study, so in total $(2^6-1)=63$ models can be obtained by considering all possible subsets of the explanatory variables. Out of these 63 possible models, for those models in which any one of the following two cases arise *viz*., a) both $X_1$ and $X_4$ and b) either $X_5$ or $X_6$, maximum likelihood estimates could not be found as the iterative procedure did not converge. There are 11 such models in which none of these two cases arise. In this study, only these 11 models have been discussed for which no problem concerning the validity of the fitted model and the existence of

MLE arose. The estimated values of parameters along with the standard error for the 11 fitted models are given in Table 2. These 11 models are denoted by M1 through M234.The model M1, whose parameter estimates were $\hat{\beta}_0 = -4.60$ and $\hat{\beta}_1 = 3.46$ can be written as

$$\pi = P(Y=1|X_1 = x_1) = \frac{1}{\left(1+e^{4.60-3.46X_1}\right)} \qquad (2)$$

Similarly the other ten models can be represented.

**Table 2.** Model, MLE of parameters and standard error of the estimates

| Model | Variable subset | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ |
|---|---|---|---|---|---|
| M1 | $X_1$ | -4.60 (0.99) | 3.46 (0.71) | - | - |
| M2 | $X_2$ | -9.22 (1.98) | 0.20 (0.04) | - | - |
| M3 | $X_3$ | -7.31 (1.46) | 7.92 (1.51) | - | - |
| M4 | $X_4$ | 0.49 (0.21) | -0.49 (0.21) | - | - |
| M12 | $X_1, X_2$ | -11.33 (3.06) | 2.88 (0.85) | 0.16 (0.06) | - |
| M13 | $X_1, X_3$ | -6.22 (2.12) | 3.20 (0.75) | 2.15 (2.29) | - |
| M23 | $X_2, X_3$ | -12.02 (2.76) | 0.18 (0.05) | 4.11 (2.01) | - |
| M24 | $X_2, X_4$ | -9.83 (2.13) | 0.22 (0.05) | 0.41 (0.37) | - |
| M34 | $X_3, X_4$ | -7.46 (1.53) | 8.11 (1.59) | -0.49 (0.29) | - |
| M123 | $X_1, X_2, X_3$ | -10.82 (3.31) | 3.00 (0.93) | 0.17 (0.06) | -1.03 (2.91) |
| M234 | $X_2, X_3, X_4$ | -12.09 (2.77) | 0.19 (0.05) | 3.84 (2.09) | 0.18 (0.39) |

Figures in brackets indicate standard errors and the numbers used in model name indicates the suffix of the independent variables

The results related to the goodness of fit and the predictive ability of the fitted models based on 108 observations has been summarized in Table 3 and 4 respectively. Perusal of Table 3 reveals that all the models except M3 are fitted well because in all the models except M3, a high probability (p) value has been observed (for testing goodness of fit, model M4 could not be considered because here in model, the regressor variable is $X_4$ which is nominal in nature resulting only in two groups, so degree of freedom for chi-square test came out to be zero).

**Table 3.** Hosmer and Lameshow (H-L) goodness of fit test

| Model@@ | Chi Square | Degrees of freedom | P-value |
|---|---|---|---|
| M1 | 0.57 | 3 | 0.90 |
| M2 | 1.99 | 8 | 0.98 |
| M3 | 9.40 | 8 | 0.31## |
| M12 | 1.92 | 7 | 0.96 |
| M13 | 1.95 | 8 | 0.98 |
| M23 | 4.32 | 8 | 0.83 |
| M24 | 1.99 | 7 | 0.96 |
| M34 | 3.55 | 8 | 0.90 |
| M123 | 1.62 | 8 | 0.99 |
| M234 | 3.84 | 8 | 0.87 |

## indicates lack of fit of model on the basis of H-L test @@ for model description, refer Table 2

Table 4 revealed that the values of Somers'D and Gamma statistics are quite high enough for all the models except M4 where the values are quite low (0.24 and 0.46 respectively). Apart from M4, the values of these two statistics for the model M3 is relatively smaller as compared to the other fitted models. The values of Somers'D and Gamma statistics are highest (0.97 for both) for the model M123. Thus the

predictive ability of all the models except model M4 is high enough and it is highest in the model M123 where comparatively more number of regressors are used indicating that more the information content, better the model.

The results of the comparison among various logistic regression models and also with discriminant function analysis in terms of classificatory power are tabulated in Table 5. A comparison was made among various logistic regression models in terms of classificatory power and it was found that all the models except M4 have high classificatory power.

**Table 4.** Predictive ability of models

| Model | Percent Concordant | Percent Discordant | Percent Tied | Somers'D | Gamma |
|---|---|---|---|---|---|
| M1 | 93.50 | 0.70 | 5.80 | 0.93 | 0.99 |
| M2 | 95.40 | 4.40 | 0.10 | 0.91 | 0.91 |
| M3 | 89.80 | 10.20 | 0.00 | 0.79 | 0.79 |
| M4 | 37.60 | 14.00 | 48.50 | 0.24 | 0.46 |
| M12 | 98.50 | 1.40 | 0.10 | 0.97 | 0.97 |
| M13 | 96.80 | 3.20 | 0.00 | 0.94 | 0.94 |
| M23 | 96.30 | 3.60 | 0.10 | 0.93 | 0.93 |
| M24 | 95.70 | 4.20 | 0.10 | 0.92 | 0.92 |
| M34 | 90.60 | 9.40 | 0.00 | 0.81 | 0.81 |
| M123 | 98.60 | 1.40 | 0.00 | 0.97 | 0.97 |
| M234 | 96.20 | 3.70 | 0.00 | 0.93 | 0.93 |

Among various fitted logistic regression models considered, the performance of the model having $X_1$ as a single explanatory variable is best in terms of classification since the values of Hit rate, Sensitivity and Specificity are high enough and the values of False positive rate and negative rate are very low.

Moreover when comparison were made between logistic regression models involving quantitative explanatory variables and discriminant function analysis, it is clear that for logistic regression model, the values of hit rate, sensitivity and specificity are higher than the corresponding value obtained from discriminant

function analysis and those of false positive and negative rate are lower than it's counterpart. Thus, it can be found that logistic regression models perform better than the alternative discriminant function analysis methods in terms of classificatory ability for the datasets considered.

**Table 5.** Comparison between results obtained from Logistic Regression (LR) modeling and Discriminant Analysis (DA) function methods on the various data sets considered

| Method | Models | Hit rate | Sensitivity | Specificity | False positive rate | False negative rate |
|---|---|---|---|---|---|---|
| LR | M1 | 92.59 | 94.12 | 90.00 | 5.88 | 10.00 |
| DA | $Z = -2.350 + 1.288*X_1$ | 88.89 | 88.24 | 90.00 | 6.25 | 18.18 |
| LR | M2 | 74.07 | 70.59 | 80.00 | 14.29 | 38.46 |
| DA | $Z = -4.018 + .078*X_2$ | 70.37 | 70.59 | 70.00 | 20.00 | 41.67 |
| LR | M3 | 92.59 | 88.24 | 100.00 | 0.00 | 16.67 |
| DA | $Z = -4.469 + 4.369*X_3$ | 88.89 | 83.33 | 100.00 | 0.00 | 25.00 |
| LR | M4** | 40.74 | 37.50 | 45.46 | 50.00 | 66.67 |
| LR | M12 | 88.89 | 88.24 | 90.00 | 6.25 | 18.18 |
| DA | $Z = -3.352 + 0.884*X_1 + 0.034*X_2$ | 81.48 | 81.25 | 81.82 | 13.33 | 25.00 |
| LR | M13 | 85.19 | 88.24 | 80.00 | 11.77 | 20.00 |
| DA | $Z = -2.564 + 1.226*X_1 + 0.319*X_3$ | 81.48 | 88.24 | 70.00 | 16.67 | 22.22 |
| LR | M23 | 81.48 | 77.78 | 80.00 | 12.50 | 38.33 |
| DA | $Z = -4.511 + 0.063*X2 + 1.208*X_3$ | 66.67 | 64.71 | 70.00 | 21.43 | 46.15 |
| LR | M24** | 81.48 | 82.35 | 80.00 | 12.50 | 27.27 |
| LR | M34** | 81.48 | 82.35 | 80.00 | 12.50 | 27.27 |
| LR | M123 | 85.19 | 88.24 | 88.89 | 6.25 | 20.00 |
| DA | $Z = -3.143 + 0.937*X_1 + 0.036*X_2 - 0.411*X_3$ | 81.48 | 82.35 | 80.00 | 12.50 | 27.27 |
| LR | M234** | 77.78 | 76.47 | 80.00 | 13.33 | 33.33 |

** corresponding to these models, discriminant function analysis has not been considered because there were qualitative variables as well

## Conclusion

In the present paper, logistic regression model has been successfully used for classification of the overall discomfort of the farm labourers during farm operation. It has been found that load given to various farm machineries has maximum influence on the discomfort level of farm labourers during farm operation. As regards to goodness of fit, all the models fitted well except when the response variable is modeled only in terms of $O_2$ consumption per minute during farm operation. Predictive ability of various fitted models is also found high enough. As compared to discriminant function analysis, the fitted logistic regression models performed well in terms of classification of new observations. Among various fitted logistic regression model considered, the performance of the model having load given to farm machineries as explanatory variable is found best for classifying discomfort level of farm labourers during farm operation.



## References

Gent, D.H., Schwartz, H.F., Ishimaru, C.A., Louws, F.J., Cramer, R.A. and Lawrence, C.B. (2003) *Phytopathology* **94,** 184 - 195.

Johnson, D.A., Alldredge, R., and Vakoch, D.L. (1996). Potato late blight forecasting models for the semiarid environment of South-Central Washington, *Phytopathology*, **86,** 480-84.

Mila, A.L., Carriquiry, A.L. and Yang, X.B. (2004) *Phytopathology* **94,** 102 - 110.

Passmore, R. and Durrnin, J. (1955) *Physiological Reviews* **35**, 801 - 875.

Ram, R., Kumar, A., Singh, A.K., Jha, S.K. and Ramasubramanian V. (2008) *J. Agri. Eng.* **45,** 12 - 18.

Saha, P.N., Dutta, S.R., Banerjee, P.K. and Narayane, G.G. (1979) *Ergonomics* **22**, 1059 - 1071.

Vergara, M. and Page, A. (2002) *Applied Ergonomics* **33,** 1 - 8.

**Running Title:** Logistic Regression for Classification in Agricultural Ergonomics

**Running Author**: Arpan Bhowmik *et al.*

Received April, 2011 : Accepted July, 2011