A logistic regression model for prediction of premonsoon convective development over Kolkata

S Dasgupta* & U K De
Atmospheric Science Research Group, Environmental Science Programme, Faculty of Science, Jadavpur University, Kolkata 700 032

Received 18 June 2003; revised 13 January 2004; accepted 27 April 2004

Certain parameters (both dynamic and thermodynamic) have been identified as significant for the occurrence of convective developments in the premonsoon period (March-May) in Kolkata (India). In the present paper, an attempt has been made to develop a logistic regression model for prediction of the risk of occurrence of these convective developments from a knowledge of the values of significant parameters. The model was developed on the basis of a sample of 87 days covering the premonsoon period of the year 1990. When used for the purpose of validation on the basis of data of an arbitrarily chosen year, it yielded about 80% correct predictions.

Keywords: Regression model; Convective development; Premonsoon prediction
PACS No: 92.60. Ek; 92.60. Ry; 92.60. We
IPC Code: GOIW 1100; GOIW 1110

1 Introduction
In recent years various statistical models have drawn the attention of scientists for the analysis of data arising from different fields.

Keeping this in view, Dasgupta and De1 have previously shown as to how a simple probabilistic description of the thunderstorm phenomenon can be provided with the help of Markov chains. The present work is an attempt to take a step ahead and set up a suitable model for the prediction of any convective development which includes noting of lightning, dry thunderstorm, thunderstorm with rain and local severe storm.

The logistic regression model2, which is based on the logistic function, is generally used to study the nature of dependence of a dichotomous response variable (Y) on a number of independent or explanatory variables X1, X2, ……Xk, which may be either discrete or continuous in nature. The logistic regression is a popular statistical modelling procedure used in the analysis of epidemiological data2.

The use of logistic regression in the context of meteorological data is of a recent origin. Sanchez et al.3 have applied this model to the short-term forecast of hail risk in the province of Leonin in the north-west Iberian peninsula of Spain.

The present work is concerned with extending this model for establishing an objective forecast of the risk of occurrence of convective development in the premonsoon period (March-May) in Kolkata (India).

2 Data and methodology
The present work is based on data relating to the occurrence of premonsoon convective developments in Kolkata in the year 1990 as noted in Alipore and Dumdam surface meteorological observatories together with the information on the significant parameters as observed from the radio sounding done regularly at Dumdam at 0000 hrs UTC and 1200 hrs UTC. It should be noted that any convective development noted in either Alipore or Dumdam is considered as convective development for Kolkata and the lead time between the observation of the predictors and the occurrence of the convective event is almost 24 hours.

Ghosh et al.4 have identified the significant parameters for the occurrence of premonsoon convective developments in Kolkata separately for the morning (0000 hrs UTC) and afternoon (1200 hrs UTC). Of these, the parameters appearing in both the morning and afternoon analyses have been identified as significant for the occurrence of premonsoon developments in Kolkata, in general, over the day.

It may be relevant to mention in this context that Ghosh et al.4 have considered the various parameters...
for the different atmospheric layers, confining the study up to the 500 hPa level. The atmosphere up to the 500 hPa level has been divided into four layers, viz. 1000-850 hPa, 850-700 hPa, 700-600 hPa and 600-500 hPa.

Consequently, the common significant parameters identified for the present study were \( \theta_e - \theta_a \) at 1000 hPa, \( P - P_{LCL} \) at 1000 hPa, \( d \theta_e /d z \) at 1000-850 hPa and \( \theta_e - \theta_a \) at 850 hPa. The parameters \( \theta_e \) and \( \theta_a \) denote, respectively, the saturated equivalent potential temperature and equivalent potential temperature; \( P \) and \( P_{LCL} \) are the pressures at the reference level and at the corresponding lifting condensation level and \( d \theta_e /d z \) is the potential convective instability of the atmospheric layer.

In fact, both \( \theta_e - \theta_a \) and \( P - P_{LCL} \) give the measure of unsaturation in the atmosphere. But the two parameters give the representation of the same quantity in two different ways. The first one is in terms of temperature units, whereas the second one is in terms of pressure units. So \( X_1 \), \( X_2 \) and \( X_3 \) are the parameters at different levels giving the degree of unsaturation.

The symbol \( d \theta_e /d z \), which represents the potential convective instability, may be rendered negative. In case it is negative, it implies that the saturation levels become gradually closer if the parcels are lifted adiabatically from the lower to upper levels.

Having identified these parameters, a suitable regression model was sought, which could be recommended for the objective forecasting of premonsoon convective developments in Kolkata.

With the \( i \)th day, \( i = 1,2,\ldots,n \), we associate a variable \( Y_i \) such that,

\[
Y_i = \begin{cases} 1, & \text{if a convective development occurs on the } i\text{th day} \\ 0, & \text{if a development does not occur on the } i\text{th day} \end{cases}
\]

Further, we take

\[
\begin{align*}
X_1 &= (\theta_e - \theta_a) \text{ at 1000 hPa} \\
X_2 &= (P - P_{LCL}) \text{ at 1000 hPa} \\
X_3 &= d \theta_e /d z \text{ at 1000-850 hPa} \\
X_4 &= (\theta_e - \theta_a) \text{ at 850 hPa}
\end{align*}
\]

It may be mentioned that the values of the covariates are obtained by averaging the 0000 hrs UTC and the 1200 hrs UTC values.

The observed values of \( X_j \) for the \( i \)th day is denoted by \( X_{ij} \).

Thus, we have a framework, where the response variable \( Y \) is dichotomous and the independent variables have no restrictions on their nature.

The general multiple linear regression model is given by

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} \\
i = 1,2,\ldots,n
\]

where each coefficient \( \beta_j \) is a measurement of the effect on the response variable of a unitary increase in the variable \( X_j \); the other explanatory variables remaining constant.

But, the function \( f(X_1,X_2,\ldots,X_n) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \) is unbounded and hence the above regression equation might provide estimates of \( Y \) which are not bounded between 0 and 1 and is thus not strictly valid here.

So, we search for a function \( f() \) which is necessarily bounded between 0 and 1.

Such a function is the logistic function given by

\[
f(X_1,X_2,\ldots,X_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}}
\]

Since the above function is bounded between 0 and 1, if we take as the regression equation

\[
Y_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}}
\]

it will ensure estimates of \( Y_i \) which are bounded between 0 and 1.

For the validity of the model in the true sense, we take, as the response variable, a probability and the explicit regression equation is taken to be

\[
\Pr (Y_i = 1 | X_1, X_2, \ldots, X_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}}
\]

or,

\[
Z_i = \Pr (Y_i = 1 | X_1, X_2, \ldots, X_i)
\]

where,

\[
Z_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}}
\]

or,

\[
\ln \frac{Z_i}{1 - Z_i} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_n X_{in}
\]
The transformation \( \ln(Z/(1-Z)) \) is called the logistic transformation and it is seen to be linear in the parameters.

Suppose in a given sample of \( n \) days, there are \( m \) days where a convective development occurs and \( n-m \) fairweather days. Then the probability of the observed sample is

\[
\frac{\Pr(Y_i = 1 | X_1, X_2, \ldots, X_6)}{\Pr(Y_i = 0 | X_1, X_2, \ldots, X_6)} = \left(1 + e^{-\beta_0 + \beta_1 X_{1i} + \ldots + \beta_6 X_{6i}}\right)^{-m}
\]

Given \( X_1, X_2, \ldots, X_6 \), the above expression is a function of \( \beta_0, \beta_1, \ldots, \beta_6 \) and is called the likelihood function of \( \theta \) where \( \theta = (\beta_0, \beta_1, \ldots, \beta_6) \) and is denoted by \( L(\theta) \).

Thus

\[
L(\theta) = \left(1 + e^{-\beta_0 + \beta_1 X_{1i} + \ldots + \beta_6 X_{6i}}\right)^{-m}
\]

The maximum likelihood estimates of \( \beta_0, \beta_1, \ldots, \beta_6 \) are obtained from the equations

\[
\delta L(\theta) / \delta \beta_i = 0 \quad \ldots \quad (1)
\]

\[
\delta L(\theta) / \delta \beta_i = 0 \quad \ldots \quad (2)
\]

\[
\delta L(\theta) / \delta \beta_i = 0 \quad (k+1) \quad \ldots \quad (3)
\]

In practice, the fitting process required by the model involves repeated numerical resolution of equations to estimate the coefficients \( \beta_i \).

### 3 Results and discussion

The sample under consideration included 87 days covering the premonsoon period (March-May) of the year 1990. Using raw data obtained from a local radio sounding and two weather stations, the data series for each of the four explanatory variables was obtained. Figure 1 shows the evolution of the four explanatory variables against the serial number of the thunderstorm days.

The response variable \( Y \) was known, given that the observation network allowed the sample of situations to be classified into thunderstorm and non-thunderstorm days.

Thus, the present data set was of the form \((Y_i, X_{1i}, X_{2i}, X_{3i}, X_{4i})\), \(i = 1(1)87\).

The coefficients of the logistic regression model were then estimated using the unconditional likelihood approach, as the number of parameters was small in comparison to the total number of observations (Table 1). Inbuilt computer programmes within the SPSS package were used for this purpose. The fitted model was obtained as

\[
\Pr(Y_i = 1 | X_1, X_2, \ldots, X_4) = \frac{1}{e^{(0.255 + 0.220 X_1 - 0.013 X_2 - 0.480 X_3 - 0.540 X_4)}}
\]

Using this model, the probability/risk of a convective development was computed for an arbitrarily chosen sample of 69 days, covering the premonsoon period of the year 1991.

In line with the works of previous researchers, the following criterion was adopted for the present study.

If \( Z_i \geq 0.5 \), there is a potential risk of occurrence of a convective development.

If \( Z_i < 0.5 \), there is practically no risk of occurrence of a convective development.

It is, therefore, reasonable to expect that on days the thunderstorms are observed, the predicted probability of occurrence should be greater than 0.5, whereas on fairweather days, the model should ideally predict a value of \( Z_i \) much less than 0.5. Actual verification gave the results as shown in Table 2.

It is evident from the results of verification that the observed and expected findings agreed on 53 out of the 69 days, which makes the overall percentage of correct predictions of 78%, a satisfactory one.
Table 1—Estimates of the coefficients of the logistic regression model

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Coefficients (β)</th>
<th>Standard error of estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant -0.2550 (β0)</td>
<td>0.4767</td>
<td></td>
</tr>
<tr>
<td>X1 0.2203 (β1)</td>
<td>0.1813</td>
<td></td>
</tr>
<tr>
<td>X2 -0.0212 (β2)</td>
<td>0.0374</td>
<td></td>
</tr>
<tr>
<td>X3 -16.4819 (β3)</td>
<td>47.1481</td>
<td></td>
</tr>
<tr>
<td>X4 -0.0492 (β4)</td>
<td>0.0264</td>
<td></td>
</tr>
</tbody>
</table>

Table 2—Results of cross-validation of the logistic regression model (Period: March - May 1991)

<table>
<thead>
<tr>
<th>Predicted situation</th>
<th>Observed situation</th>
<th>Convective development</th>
<th>Fairweather</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convective development</td>
<td>29</td>
<td>10</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Fairweather</td>
<td>6</td>
<td>24</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>34</td>
<td>69</td>
<td></td>
</tr>
</tbody>
</table>

Note:
Success rate: 29/35 = 0.8285714
False alarm rate: 10/34 = 0.294176
Observed value of the chi-square statistic: 0.9435897
The hypothesis of goodness of fit has been accepted.

However, it should be borne in mind that the size of the sample used for the purpose of setting up the model is not very large. Further, interaction effects between the different explanatory variables have not been considered for the present study. Besides, a few more parameters (both dynamic and thermodynamic) could have been included as explanatory variables. With these modifications, the fit of the model is expected to be even more satisfactory and the overall percentage of correct predictions, when it comes to cross-validation of the model will, perhaps, be further increased.

4 Conclusions
Despite the limitations of the present study, it clearly stands out that the logistic regression model, once set up, is easy to use and the speed of the calculation makes it an extremely handy forecasting tool for the prediction of the risk of premonsoon convective developments in Kolkata. All that are required are a local radio sounding and certain meteorological data from a weather station at least.

It may be relevant to add that, although examined in the light of a local atmospheric phenomenon, this model can, perhaps, be used whenever one is concerned with the prediction of the chance of occurrence of any atmospheric phenomenon anywhere in the world. In other words, the statistical regularity that is observed locally may be observed globally.

Acknowledgements
The authors wish to thank the India Meteorological Department for the supply of data and valuable references. The authors also wish to thank Mr Piyush Basak and Mr Subhra Chatterjee for making the data available to them in the form as desired by the model. One of the authors (S D) is personally indebted to Mr. Lalit Talwar for making available the latest version of the SPSS package and to Prof. Ayan Chandra of St. Xavier’s College, Kolkata, for providing valuable help in running the programme.

References