PREFACE

In order to establish a prediction model for bird-migration, we used the method of multiple linear regression. The first point of this paper provides a brief summary to theoritical considerations on the multiple regression model, to make more easily the understanding of the paper, as to notation and vocabulary.

On the last meeting of the Bird/Radar/Weather Working Group, Mr. Houghton said that all workers probably needed more experience in the use of different kinds of multiple regression to develop the best method for successful bird movement forecast.

Therefore, the rest of the paper attempts to give an answer to the following questions.

1. Can we reduce the space of variables without loosing information, how can we automatically eliminate redundant variables ?
2. How do we consider the correlation bird-movement versus meteo ?
3. Which methods of multiple regression can we use ?
4. Can we use the multiple regression model as a prediction tool ?

It would be very interesting to know if someone else would try to use the methods explained in this paper with their data. I can provide information as to the employed programs (written in FORTRAN IV and used on the IBM 360/65) as to the theoritical considerations.

However it is possible that bird migration is dependent on "weather" without the possibility of explaining it in a linear regression form. The aim is to find this dependence. It is not excluded that with this information, other methods (of which Mr. Louette is reporting) become valuable for prediction.

I. THE MULTIPLE LINEAR REGRESSION MODEL

A. We dispose of n observations of p variables

$Y, X_1, X_2, \ldots X_{p-1}$

Y is called dependent or "explained" variable or regressor

X_i $i = 1, \ldots, p-1$ independent variable, or explanatory or regressand

The theoretical "suspected" relation is :

$$y_{(n,1)} \quad \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$X_{(n,p)} \quad \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{pmatrix}$$

$$\beta_{(p,1)} \quad \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

$$E_{(n,1)} \quad \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\text{or} \quad \forall i: \quad y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \varepsilon_i \qquad (1)$$

$$y = X\beta + E$$

The relations of y are = observed x . β (which are unknown) + a stochastic vector (unknown). So we have to determine the unknown values of β by making the following assumptions on E

$$E_{(n,1)}(E) = 0 \qquad (\text{for all } i \quad E(\varepsilon_i) = 0)$$

$$V(E) = \sigma^2 I_n$$

B. Estimation of the parameters ($\hat{}$ = estimation)

$$\hat{\beta} = (x'x)^{-1} x'y$$

$$E(\hat{\beta}) = \beta$$

$$V(\hat{\beta}) = \sigma^2 (x'x)^{-1}$$

If we assume that :

$$E \rightsquigarrow \mathcal{N}_n (0, \sigma^2 I_n)$$

than :

$$\hat{\beta} \rightsquigarrow \mathcal{N}_p (\beta, \sigma^2 (x'x)^{-1})$$

$$\hat{\sigma}^2 \rightsquigarrow \frac{\sigma^2}{n-p} \chi^2_{n-p}$$

the variables $\hat{\beta}$ and $\hat{\sigma}^2$ are indpendent in probability

R^2 may be viewed as measuring the relative amount of variation in the dependent variable that is "explained" by the independent variable.

$$R^2 = \frac{\hat{\beta}' x'y - n\bar{y}^2}{y'y - n\bar{y}^2}$$

C. We also can standardise the variables

$z_p = (y - m_p)/(s_p)$ and $z_j = (x_{ij} - m_j)/s_j$

with $m_p = \sum y_i / n$

$m_j = \sum x_{ij} / n$ $\qquad j = 1, \ldots, p$

$s_p = \sum (y_i - m_p)^2 / n$

$s_j = \sum (x_{ij} - m_j)^2 / n$

We can formulate the same model. The estimators become (by introducing the standardized variable z_j in the relation (1) and with

$$\hat{\beta}_j = \frac{s_y}{s_j}\hat{\beta}'_j \qquad j = 1, \ldots, p$$

$$\hat{\beta}_0 = \left(\bar{y} - m_1 \hat{\beta}'_1 - \ldots - m_p \hat{\beta}'_p\right)/s_y$$

$$\hat{\sigma}^2 = \hat{\sigma}'^2/s_y^2$$

and if calling R the matrix of empirical correlation of the p variables taken into consideration, we can write R as

$$R\ (p \times p) = \begin{pmatrix} R_{11}(p-1 \times p-1) & R_{12}\ (p-1 \times 1) \\ R_{21}(1 \times p-1) & R_{22}\ (1 \times 1) \end{pmatrix}$$

We have to estimate

$$\hat{\beta}' = \begin{pmatrix} \hat{\beta}'_0 \\ \hat{\beta}'_{p-1} \end{pmatrix} = \frac{1}{\alpha} \begin{pmatrix} r \\ R_{11}^{-1} r_{12} \end{pmatrix}$$

$$V\hat{\beta}' = \hat{\sigma}'^2 \begin{pmatrix} * & * \\ * & R_{11}^{-1} \end{pmatrix}$$

$$\hat{\sigma}'^2 = \frac{n}{n-p}\left(1 - R_{12} R_{11}^{-1} r_{12}\right)$$

$$V\hat{\beta}' = \hat{\sigma}'^2/n-p$$

$$\alpha^2 = 1 - R_{12} R_{11}^{-1} R_{12}$$

We decided the use of standardized variables.
It's possible to switch from variable in standardised form to the original form.

D. It's also possible to appreciate the validity of the numeric results of the regression.
Several tests of signification are generally used.

(1) test of signification for 1 variable :

H_0 : $\alpha_j = 0$ $\forall$ j

against H_1 : $\alpha_j \neq 0$

(2) test of signification for a group of var iables

H_0 : $\alpha_j = 0$ for $j = \ldots$

H_1 : $\alpha_j \neq 0$ for $j = \ldots$

(3) A total test of signification for the regression model

H_0 : $\alpha_j = 0$ $\quad j = 0, 1, \ldots, p-1$

H_1 : $\alpha_j \neq 0$ $\quad j = 0, 1, \ldots p-1$

R^2 is a measure to test the validity of the regression, but it is obvious that there are still other criteria, certainly when using that method as a prediction tool. R^2 is a non decreasing function of the number of variables used. The difference between Y observed and Y estimated cannot be too large etc.

E. One can find out immediatly which problems arise à priori for the user of this method.

Which variables do we have to choose as independent variable and as dependent variable ?

(1) either both are fixed (instruments)
(2) either the dependent variable is fixed and one has to determine the measure of the independent variables and vice versa
(3) either both types of variables are to be choosen

As dependent variable (here : the bird migration intensity) we may take :

- the intensity average by day or by night or at a fixed time
- the max-intensity or the min intensity
- the total intensity by day and night etc.

As independent variable (here : a meteo-variable)
- at what time ?
- the locality where the variable had been measured ?
  etc.

The choice of the dependent variable in each practical case is let to the "intuition" of the user.
As to the independent variables to be introduced we try to give an answer in this paper.

II. WHAT KIND OF RELATION DO WE INTEND TO STUDY ?

Do we want an answer on such a question as :
Is there any relation between the bird migration intensity and weather (defined by the meteo data available) ?
Or do we want an answer on the question.
Is there any relation between the bird migration intensity and some of the meteo variables ?

1. Is there any relation between the bird migration intensity and weather ?

Under "weather total" we understand : all variables furnished by the meteo-stations. In our case 23 variables were available. The treatment of such a number of variables is always complex. Is it not possible to reduce that number of variables without loosing information ? Can we eliminate some variables à priori (correlated, redundant, and variables with a very low contribution to the information "weather"). In other words we'll analyse the data reducing the space of the variables and loosing a minimum of information.
The evolution and the variety of the technics used to analyse data, is due indoubtly to the evolution of the computer, an indisputable tool to apply this methods. The method we applied is known as Factorial analysis and analysis in principal components.
This method is explained in a lot of technometric, psychometric and statistic ~~periodics~~ references. I'll only mention the basic principals of the method without getting into details.

Assume we have n observations of p variables.
Generally we have to handle with such dimensions (p or n) that they are an obstacle for the user to assimilate the information So we intend (as far as possible) to reduce the dimensions with a minimum less of information. May we expect to get down to 2, 3 or 4 dimensions ?

The distance between two points is taken as caracteristic measure in the space of the variables (dimension p) (working in the space of observation (dimension n) is exactly the same). One wants to project the distances on a space with dimension $q << p$ such that the distances between the projected points and those of the space of dimension p does not differ much. One likes having the lenght of projections in average maximal, such that the distortion of the cloud of points in the space of dimension p is reduced to a minimum.

Note that it is advisable to use standardized variables because if the distance $(x_{1j} - x_{2j})^2$ from the j^{th} variable is larger than the weight of this term in the sum $\sum_{i=1}^{p} (x_{1i} - x_{2i})^2$ would be too important.

Taking into account the chosen criterium and using standardizet variables we have to handle the correlation matrix R in that way that we get a matrix L of the eigen vectors corresponding to the eigen values of R.

The q^{th} vector will be the q^{th} principal component and corresponds with the q^{th} eigen value (in decreasing order).

The quality of that information can be measured by the cumulative percentage of eigen values

$$\sum_{i=1}^{q} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$$

To get finally a matrix more interpretable than the matrix L we used Kaiser's method cfr ref (7). We survey the methods of multiple regression here after on the selected variables.

2. Is there any relation between the bird migration intensity and SOME of the meteo variables ?

Between the variables "suspected" to explain Y we have to select a minimum number of variables which are sufficient (in the sense of theory of test) to explain the variations of Y.

Some generalities

Let p be the number of variables suspected to explain y.
Let J be the set of the integers 1, ... p and let J_k be a subset of J with cardinal k.

Let CCM(.) be the function, associated with a subset of variables J_k, the multiple correlation coefficient of y on these variables. The best regression with k variables, noted J_k^*, is that regression for which the multiple regression coefficient between the C_p^k regressions is a maximum.

$$\forall\ J_k \quad CCM\ (J_k^*) \geqslant CCM\ (J_k)$$

the problem :

The confidence level α being fixed (1) for a given k ($k = 0, 1, \ldots, p$) to determine the subset J_k^* resulting into the regression $R_k^{2*} = CCM(J_k^*)$

(2) to determine the minimum value of k, say k^*, which allows us to accept the total non-significance of the variables of the set $J - J_k^*$

The best regression on the level α is than defined by $J_{k^*}^*$

Note that :

$$0 \leq R_0^2 \leq R_1^{2*} \leq \ldots \leq R_{k-1}^{2*} \leq R_k^{2*} \leq \ldots \leq R^2 \leq 1$$

and that the best regression with k-1 variables is not necessary contained in the best regression with k variables.

$$J_{k-1}^* \not\subset J_k^*$$

Mathematical formulation :

a. for k = 0 to p (step 1) to determine J_k^* such that

$$\max_{J_k} CCM(J_k)$$

b. α fixed : to determine the smallest value of k which satisfies the level α

The method based on "Best subset search" Algorithm AS 38 (cfr ref) of GARSIDE, conducts to the exact solution. It's also possible to eliminate the redundant variables by the method described in [1].

Backward regression and stepwise regression allows us to get approximate solutions.

Remarks :

We applied all these methods on one data.
We used the method described in II 1. to reduce the space of our meteo-variables
autumn 71 - spring 71 - autumn 72 - autumn 71
the whole year 71 and the whole year 72.

We could see that always the same variables were selected, sometimes in different order but at least we kept 90 % of the total information. We used the variables selected by this method as independent variables. As dependent variable we studied the following problem :

What can we use as a good measure of migration ?
We establish a program called SUNSET-SUNRISE which gives us the fluctuation of the counted echo's during day and night (e.g. the average number of echo's per observation during a given period in the classes sunset, hour after Sunset or Sunrise + 1 to + 2 - 1 to - 2 etc.

As dependent variable we took the square root of the average number of echo's per observation during a given period of day or/ and night depending on the period in the year.

We used all the regression methods described in this paper. But regression is not all, we want to have a prediction tool. We have a need for bringing about certain desired results in the future. One rational way to achieve or nearly to achieve this goal is to study the relevant variables in the behavior of interest and to project the pattern of behavior on the assumption that the pattern that has persisted will hold true, in the future, so the esimates of the parameters of the model have to be used for making a prediction. So we are interested to know if we'll find the same results for the same periods in 71 and 72.

If we want to have a good prediction tool we believe that we have to use few variables and we have to obtain regression coefficients mostly the same in each autumn period and spring period.

If a good prediction tool is not available we are sure to obtain a classification of variables which have an influence on the bird migration.

Conclusions :

In the time this paper is written we do'nt have all results. Probably we 'll be ready before this meeting such that we are able to illustrate the importance of this methods explained in this paper.

BIBLIOGRAPHY

1. GARSIDE J. (1971) "Some computational procedures for the best subset problem" Applied Statistics,Vol.20,Nr.1.

2. LEBART L. et FENELON J-P. (1971) "Statistique et informatique appliquées" Ed. Dunod,Paris.

3. KAISER (1959) "Computer program for varimax rotation in factor analysis" Educational and Psychological measurement,Vol.XIX,Nr.3.

4. SAUTEREAU C.(1972) "Sélection des variables à faire intervenir dans une régression linéaire" Rapport de recherche SVR/1 - 8/72 CIRO.

5. SCHATROFF M. (1968) "Efficient calculation of all possible regressions" Technometrics,Vol.10,Nr.4.

6. ULMO J. (1971) "Problèmes et programmes de régression" Revue de statistique appliquée,Vol.19,Nr.1.

7. "Best subset search" Algorithm AS 38. Applied Statistics,Vol.20,Nr.1.