Journal of Animal and Veterinary Advances 5(6) : 522-527, 2006

Predicting the Probability of Conception in Artificially Inseminated Bovines-A Logistic Regression Analysis

M. Thirunavukkarasu and G. Kathiravan
Department of Animal Husbandry Statistics and Computer Applications,
Madras Veterinary College, Chennai-600 007 India

Abstract: A study was undertaken to estimate the probability of conception in artificially bred bovines, based on various animal and environmental factors, based on the data collected from 2,283 bovines (1,942 cattle and 241 buffaloes) inseminated at 30 artificial insemination centres in six districts of North-eastern agroclimatic zone of Tamil Nadu State (India). Logistic regression technique was employed to estimate the probability of a particular breedable bovine female not being able to conceive of an artificial insemination. Wald statistic obtained for the independent variables indicated that the coefficients for the variables-species, lactation order, stage of lactation, milk yield, reproductive disorders, distance to artificial insemination centre and month of insemination were all significantly different from '0' at 1 degree of freedom. Positive values of 'R' statistic obtained for species, stage of lactation, reproductive disorder, distance to artificial insemination centre and month of insemination indicated that as these variables increase in value, the likelihood of conception increased by 7.2795, 2.7478, 2.5638, 2.7453 and 1.9778, respectively. The negative 'R' in the case of average milk yield indicated that the likelihood of conception decreased by 0.1973. Small values of 'R' for the statistically insignificant variables such as breed, farm size, land holding and lactation order indicated that these variables had got only smaller partial contributions to the model. Again, 1510 animals not conceived of AI were correctly predicted by the model not to have conceived. Similarly, 745 animals conceived were correctly predicted to have conceived. Of the animals not conceived, 98.82% were correctly classified, while of the animals conceived, 98.68% were correctly classified. Overall, 98.77% of the 2283 bovines were correctly classified to have conceived/not conceived.

Key words: Bovines, conception, probability, logistic regression

INTRODUCTION

Artificial breeding of bovines was introduced some four decades back as a crucial strategy to ensure productivity enhancement of Indian cattle and buffaloes so as to augment milk production in the country. As a result, milk production in the country has tremendously increased from just 17 million tones in 1950's to about 85 million tones today to make India the topper in milk production across the world. However, low conception rates achieved among the artificially inseminated bovines across the country had been a bottleneck experienced in the way of promoting and sustaining the higher level of milk output. Female fertility is regulated by an animal's genetic disposition and various environmental factors[1]. These factors might include climate, management, inseminator's skills, quality of semen, animal's reproductive status, milk yield, number of days post partum, etc. Fertility of dairy cattle receives much attention because poor reproduction is costly for dairy producers, through higher breeding and veterinary fees[2,3] found many factors influencing reproductive performance of lactating dairy cows. Management factors such as accuracy of heat detection, use of proper inseminating techniques, proper semen handling and appropriate herd health management can directly influence reproductive performance of the dairy herd. In addition, some factors, which are beyond the control of management such as milk yield, age and season might also affect fertility. Although these factors are less controllable by the management, an understanding of their relationship to fertility is important for proper interpretation of reproductive ability of a herd.

A decreased conception rate, which is a result of various causes operating in the continuum between the semen production station and the insemination centre, has been found to be a negative input or disincentive both at farm and at national levels. This is especially true when genetically high potential stock is unable to deliver goods because of conception failures. A failed insemination not only lets a dairy farm to lose its

Corresponding Author: M. Thirunavukkarasu, Department of Animal Husbandry Statistics and Computer Applications, Madras Veterinary College, Chennai-600 007 India

revenue/profit, but also becomes a burden on the government exchequer through causing an 'avoidable' wastage of scarce resources in terms of employees' salaries, expenses on inputs, processing, transport and storage charges, etc. there is little documented evidence available on this front to know what is really lost, if a cow or a buffalo in oestrum fails to conceive. For this reason, this study was undertaken to estimate the probability of conception in artificially bred bovines, based on various animal and environmental factors which are believed to have a say on the conception rates achieved.

MATERIALS AND METHODS

To achieve the objective of the study, North-eastern agroclimatic zone of the State (out of the seven existing in the State), which consisted of six districts, was randomly selected, from which 30 insemination centres were selected again randomly. A total number of 2,283 bovines inseminated at these centres (1,942 cattle and 241 buffaloes) were selected at random (Table 1) and considered for the analysis. Out of the total 2,283 animals sampled, cows formed a greater proportion (85.06%) than buffaloes (14.94%).

Table 1: Number of animals studied north eastern zone of Tamil Nadu

Districts	Sampled AI Centres	No. of Cattle	No. of Buffaloes	Total
Cuddalore	6	378	26	404
Villupuram	5	421	9	430
Vellore	5	272	53	325
Kancheepuram	5	307	86	393
Thiruvallur	6	407	131	538
Thiruvannamalai	3	157	36	193
Zone total	30	1,942 (85.06)	341 (14.94)	2,283 (100.00)

(Figures in parentheses indicate percentages to total)

Logistic regression analysis: Predicting whether an event will occur or not is important in biological sciences. A variety of multivariate statistical techniques are used to predict a binary dependent variable from a set of independent variables. The most popular technique, multiple regression analysis, poses difficulties when the dependent variable can have only two values-an event occurring or not occurring. Because, the assumptions necessary for hypothesis testing in regression analysis are necessarily violated, if the dependent variable can have only two values. Further, with multiple regression analysis, the predicted values cannot be interpreted as probabilities. However, the logistic regression model was found to estimate the probability of occurrence of an event, better than other tools[4].

The logistic regression model was the technique of choice for analysing binary response variable in veterinary or human epidemiology[5,6] applied a logistic regression model for analysing the conception status of cows. Lassauzet *et al.*,[7] modeled the probability of bovine leucosis, using logistic regression, as a function of prevalence of infection in pen, presence of lactating cows, proportion of pregnant cows and presence of an infected bull. Correa *et al.*,[8] used logistic regression to model the risk factors for downer cow syndrome from the records of 2705 lactations from 12 Holstein dairy herds. Jensen and Hoier[9] derived a logistic regression model, which classified correctly 38 cases (84%) suffering from hepatobiliary diseases of 45 dogs analysed. Thirunavukkarasu and Prabaharan[10] used logistic regression model to estimate the probability of milch animal to pick up mastitis from the data on 301 animals and 148 non-mastitic animals and found that 91.98% of 449 observations were correctly predicted. Selvam[11] used logistic regression model to estimate the probability of poultry layers picking up diseases such as Ranikhet and Infectious Bursal Diseases.

In the present study, logistic regression technique was employed to estimate the probability of a particular breedable bovine female not being able to conceive of an artificial insemination. The logistic regression model set for this purpose is as below:

Prob (event) or P_i $=E(Y = 1/V_i)$

$$= \frac{e^{\delta+\gamma_1 v_1+\gamma_2 v_2+\ldots+\gamma_i v_i}}{1+e^{\delta+\gamma_1 v_1+\gamma_2 v_2+\ldots+\gamma_i v_i}} \quad i = 1, 2, 3, \ldots, 11$$

or equivalently

$$P_i = \frac{1}{1+e^{-(\delta+\gamma_1 v_1+\gamma_2 v_2+\ldots\ldots+\gamma_i v_i)}}$$

or simply

$$P_i = \frac{1}{1+e^{-Z}}$$

where,

δ and γ_i = the coefficients to be estimated from the data;

V_1= Species dummy (= 1 if cow; = 0 if buffalo);

V_2 = Breed dummy (= 1 if exotic pure/ crossbred/ Murrah/ graded; = 0 if otherwise);

V_3 = Farm size dummy (= 1 if small/large; = 0 if otherwise);

V_4 = Land holding dummy (= 1 if small; = 0 if otherwise);

V_5 = Lactation order dummy (= 1 if heifer/ 3 or less; = 0 if otherwise);

V_6= Stage of lactation dummy (= 1 if second and third; = 0 if otherwise);
V_7= Average milk yield in litres;
V_8= Presence of reproductive disorder (= 1 if no; = 0 if yes);
V_9= Distance from AI center (= 1if less than 3 km; = 0 if otherwise);
V_{10}= Month of insemination (= 1 if July to February; = 0 if otherwise);
e= the base of the natural logarithms, approximately 2.718
Z= the linear combination such that $Z = \delta + \gamma_1 V_1 + \gamma_2 V_2 + \ldots\ldots + \gamma_i V_i$

The probability of the event not occurring is estimated as

Prob (no event) = 1 - Prob (event)

The probability estimates would always be between 0 and 1, regardless of the value of Z. The logistic coefficients were estimated using the maximum likelihood technique, where the coefficients that make the observed results most `likely' were selected. To test the hypotheses about the coefficients, Wald statistic, which has a χ^2 distribution, is used in the logistic models. Wald statistic is the square of the ratio of the coefficient to the standard error.

$$\text{That is, Wald} = \left[\frac{\gamma_i}{SE_i}\right]^2$$

In logistic regression, the contribution of each variable depends on the other variables in the model. A statistic that is used to look at the partial correlation between the dependent variable and each of the independent variables is the R statistic, which could range from -1 to +1. A positive `R' value indicates that the likelihood of the event occurring increases as the variable increases in value. If it is negative, opposite is true. Small `R' values indicate small partial contribution to the model.

The Eq for the statistic is:

$$R = \pm\sqrt{\frac{\text{Wald statistic} - 2K}{-2LL(0)}}$$

where,
K is the degrees of freedom for the variable;
–2LL(0) is –2 times the log likelihood of the base model that contains only the intercept.
If the Wald statistic is less than 2K, then R is set to 0.

A measure of how well the model fits is the goodness of fit statistic, which compares the observed probabilities to those predicted by the model. This statistic is defined as:

$$Z^2 = \Sigma\frac{\text{Residual}_i^2}{P_i(1-P_i)}$$

where the residual is the difference between the observed value Y_j and the predicted value P_i.

Cox and Snell R^2 and Nagelkerke $\breve{R}^2$ are the statistics that attempt to quantify the proportion of explained variation in the logistic regression model. They are similar to the R^2 in a linear regression model, although evaluation in a logistic model needs to be defined differently. The Cox and Snell R^2 is:

$$R^2 = 1 - \left[\frac{L(O)}{L(B)}\right]^{2/N}$$

where,
L(0) is the likelihood for the model with only a constant;
L(B) is the likelihood for the model under consideration;
N is the sample size.

Nagelkerke (1991) proposed a modification of the Cox and Snell R^2 so that value of 1 could be achieved. Nagelkerke $\breve{R}^2$ is:

$$\breve{R}^2 = \frac{R^2}{R^2_{max}}$$

where,

$$R^2_{max} = 1 - [L(O)]^{2/N}$$

For interpreting the logistic coefficient, the logistic model is rewritten in terms of the odds of an event occurring. The odds of the event occurring are defined as the ratio of the probability that it will occur to the probability that it will not. The logistic model rewritten in terms of the log of the odds, called as a logit, is as follows:

$$\text{Log}\left[\frac{P_i}{1-P_i}\right] = \delta + \gamma_1 v_1 + \gamma_2 v_2 + \ldots + \gamma_i v_i$$

where the logistic coefficient (γ_i) can be interpreted as the change in the log odds associated with one unit change in the independent variable.

The logistic Eq can also be written in terms of odds, instead of log odds, for easier interpretation, as below:

$$\left[\frac{P_i}{1-P_i}\right] = e^{\delta+\gamma_1 v_1+\gamma_2 v_2+\ldots+\gamma_i v_i}$$
$$= e^{\delta+\gamma_1 v_1+\gamma_2 v_2+\ldots+\gamma_i v_i}$$

where `e' is raised to the power `γ_i' is the factor by which the odds change when the i^{th} independent variable increases by one unit.

The logistic regression analysis was carried out using the software - SPSS Regression Models™ 9.0. and

RESULTS AND DISCUSSION

The results of the logistic regression function model fitted are presented in Table 2. As it could be seen from the Table, Wald statistic obtained for the independent variables indicated that the coefficients for the variables-species, lactation order, stage of lactation, milk yield, reproductive disorders, distance to artificial insemination centre and month of insemination were all significantly different from '0' at 1 degree of freedom. The Table also presents the estimated coefficients of the independent variables incorporated in the logistic model along with their standard errors, degrees of freedom, 'R' statistic and their exponential values.

As the contribution of individual independent variables to the dependent variable in the logistic model cannot be determined, 'R' statistic was calculated based on the formula given in the Methodology part to look out for the partial correlation between the dependent variable and each of the independent variables, which would range from –1 to +1. The sign of the corresponding coefficients is attached to 'R'. The positive values of 'R' obtained for species, stage of lactation, reproductive disorder, distance to artificial insemination centre and month of insemination indicated that as these variables increase in value, the likelihood of conception increased by 7.2795, 2.7478, 2.5638, 2.7453 and 1.9778, respectively. The negative 'R' in the case of average milk yield indicated that the likelihood of conception decreased by 0.1973. Small values of 'R' for the statistically insignificant variables such as breed, farm size, land holding and lactation order indicated that these variables had got only smaller partial contributions to the model.

The logit equation indicates that the logistic coefficients can be interpreted as the change in the log odds associated with one unit change in the independent variables. As detailed in the methodology of the study, rearrangement of the logistic equation obtained in terms of the odds of event occurring is essential for interpreting the logistic regression coefficients estimated. The logit, the logistic model estimated in terms of the log of the odds, is:

$$\text{Log}\left[\frac{\text{Prob (conceived}}{\text{Prob (not conceived)}}\right] = \text{Log}\left[\frac{P_i}{1-P_i}\right]$$

$$= -6.4177 + 7.2795V_1{**} + 1.0876V_2 + 0.4904V_3 + 0.8797V_4 + 0.5593V_5 + 2.7478V_6{**} - 0.1973V_7{*} + 2.5638V_8{**} + 2.7453V_9{**} + 1.9778V_{10}{**}$$

That is, when a cattle is inseminated, with the values of other independent variables remaining the same, the log odds of the bovine getting conceived increased by 7.2795. Similarly, when the other variables such as stage of lactation, presence of reproductive disorder, distance to AI centre and month of insemination increased by one unit, *Ceteris paribus*, the log odds of the female bovine to get conceived increased by 2.7478, 2.5638, 2.7453 and 1.9778, respectively. However, when the average milk yield increased by one unit, the log odds of the animal getting conceived decreased by 0.1973.

Since, it is easier to think off odds rather than log odds, the logistic Eq can be written in terms of odds as:

$$\frac{P_i}{1-P_i} = e^{-6.4177+7.2795V1{**}+1.0876V2+0.4904V3+0.8797V4+0.5593V5+2.7478V6{**}-0.1973V7{*}+2.5638V8{**}+2.7453V9{**}+1.9778V10{**}}$$

As indicated earlier in Chapter III, the `e' raised to the power $\alpha_i$ is the factor by which the odds change when the `i^{th} independent variable increases by one unit. If α_i is positive, this factor will be greater than 1 which means that the odds are increased. If α_i is zero, the factor equals 1 which leaves the odds unchanged.

Calculating the odds of getting conceived for a sample animal in the study (a cattle, non-descript, small farm size, small holding, 2nd lactation, 3rd stage of lactation, 4 litres of milk yield, no reproductive disorder, less than 3 km from AI centre and inseminated in September month) needs the computation of the probability of the bovine getting conceived as follows:

$$\text{Estimated } P_i = \frac{1}{1+e^{-Z}}$$

Where, Z = $-6.4177 + 7.2795(1) + 1.0876(0) + 0.4904(1) + 0.8797(1) + 0.5593(1) + 2.7478(1) - 0.1973(4) + 2.5638(0) + 2.7453(1) + 1.9778(1)$ and
= 9.4765

Table 2: Parameter estimates of logistic regression model

Variable	Estimated coefficient	Standard error	Wald statistic	R statistic	Exp (B)
Species dummy	7.2795	0.5295	188.9889**	0.2540	1450.2999
Breed dummy	1.0876	0.6239	3.0390	0.0189	2.9672
Farm size dummy	0.4904	0.2676	3.3592	0.0217	1.6330
Land holding dummy	0.8797	0.7324	1.4427	0.0000	2.4102
Lactation order dummy	0.5593	0.7320	0.5837	0.0000	1.7494
Stage of lactation dummy	2.7478	0.7763	12.5296**	0.0603	15.6079
Average milk yield in litres	-0.1973	0.0810	5.9299*	-0.0368	0.8209
Presence of reproductive disorder	2.5638	0.5537	21.4421**	0.0819	12.9847
Distance from AI center	2.7453	0.5769	22.6427**	0.0844	15.5690
Month of insemination	1.9778	0.5746	11.8466**	0.0583	7.2269
Constant	-6.4177	0.7543	72.3923		

* Significant at 5% level of probability, ** Significant at 1% level of probability, (Degree of freedom for each variable is 1)

Table 3: Comparison of predictions of logistic regression to the observed outcome (confusion matrix)

Categories	Predicted	Observed	Percent correct
Non-conceived	1528	1510	98.82
Conceived	755	745	98.68
Overall	2283	2255	98.77

$$\text{Thus, } P_i = \frac{1}{1+e^{-9.4765}} = 0.999923$$

The estimated probability of the chosen animal getting conceived is therefore 0.999923 and it shows that this particular animal has very high probability of getting conceived. The probability of getting not conceived is only 0.000077 (1-0.999923).

One way to assess how well this model fits is to compare the model's predictions with the observations. Table 3 is the classification table that compares the model's predictions with the observations.

From the table, it could be seen that 1510 animals not conceived of AI were correctly predicted by the model not to have conceived. Similarly, 745 animals conceived were correctly predicted to have conceived. However, 18 animals conceived and 10 animals not conceived were incorrectly classified. Of the animals not conceived, 98.82% were correctly classified, while of the animals conceived, 98.68% were correctly classified. Overall, 98.77% of the 2283 bovines were correctly classified to have conceived/ not conceived.

Another way of assessing the goodness of fit of the model is to examine how `likely' the sample results actually are, given the parameter estimates. The probability of the observed results is known as likelihood. Since the likelihood is a small number, it is customary to use-2 times the log of the likelihood (-2LL) as a measure of how well the estimated model fits the data. A good model is one that results in a likelihood of the observed results. This translates to a small value for-2LL. If a model fits perfectly, the likelihood is 0. For the logistic regression model that contained only the constant, -2LL was 2897.934 (Table 4).

Table 4: Goodness of fit

Particulars	Value
-2LL when constant alone was included in model	2897.934
-2LL when all the independent variables were incorporated	211.045
Goodness of fit	1741.116
χ^2 of the model	2686.885**
Cox and Snell R^2	0.692
Nagelkerke $\check{R}^2$	0.962

** Significant at 1% level of probability
Degrees of freedom - 10

For the model with all the independent variables, the value of-2LL was 211.045, which is smaller than -2LL for the model containing only a constant. The goodness of fit statistic was 1741.116, while the model χ^2 which is the difference between-2LL for the model with only a constant and -2LL for the model with all independent variables was 2686.885 (2897.934-211.045) which tests the null hypothesis that the coefficients for all the terms included in the model except the constant were `0'. The degrees of freedom were the difference between the number of parameters in the two models. The values of Cox and Snell R^2 and Nagelkerke $\check{R}^2$ estimated were 0.692 and 0.962, respectively, indicating that explained model was a good fit.

CONCLUSION

The logistic regression model fitted could predict excellently well the probability of an animal getting conceived of an artificial insemination, based on the animal and environmental factors and hence this equation can again be applied in field conditions and farmers advised accordingly, before the loss due to conception failures in farmers' fields rise to alarming levels.

ACKNOWLEDGEMENT

The authors are thankful to the Indian Council of Agricultural Research for having funded this study.

REFERENCES

1. Amir, D., M. Bar-El, D. Kalay and H. Schindler, 1982. The contribution of bulls and cows to the seasonal differences in the fertility of dairy cattle in Israel, Anim. Reproduct. Sci., 5: 93.
2. Faust, M.A., B.T. Mcdaniel, O.W. Robison and J.H. Britt, 1988. Environmental and yield effects on reproduction in primiparous Holsteins, J. Dairy Sci., 71: 3092-3099.
3. Hillers, J.K., P.L. Senger, R.L. Dartington and W.N. Fleming, 1984. Effects of production, season, age of cow, days dry and days in milk on conception to first service in large commercial dairy herds, J. Dairy Sci., 67: 861.
4. Hosmer, D.W. and S. Lemeshow, 1989. Applied Logistic Regression, New York: John Wiley and Sons.
5. Dyke, G.V. and H.D. Patterson, 1952. Analysis of factorial arrangements when the data are proportions, Biometrics, 8: 1-12.
6. Ron, M., R. Bar-Anan and G.R. Wiggans, 1984. Factors affecting conception rate of Israeli Holstein cattle, J. Dairy Sci., 67: 854-860
7. Lassauzet, M.L.G., M.C. Thurmond, W.O. Johnson, F. Stewens and J.P. Picanso, 1991. Factors associated with transmission of bovine leukemia virus by contact in cows on a California dairy, Am. J. Epidemiol., 13: 164-176.
8. Correa, M.T., H.N. Erb and J.M. Scarlet, 1993. Risk factors for downer cow syndrome, J. Dairy Sci., 76: 3460-3463.
9. Jensen, A.L. and R. Hoier, 1993. Clinical chemical diagnosis of diseases assisted by logistic regression illustrated by diagnosis of canine primary and secondary hepatobiliary diseases, J. Vet. Med., 40: 102-110.
10. Thirunavukkarasu, M. and R. Prabaharan, 1998. Estimating the Probability of a Milch Animal Going Mastitic-A Logistic Approach, II International Conference on Operations and Quantitative Management, Ahmedabad.
11. Selvam, S., 2000. A study on economic implications of diseases in layer farms, Unpublished Ph.D. thesis submitted to Tamil Nadu Vet. Anim. Sci. University, Chennai (India).
12. Nagelkerke, N.J.D., 1991. A note on general definition of the coefficient of determination, Biometrika, 78: 691-692.