Journal of Crop and Weed, 14(1): 117-121 (2018)

Designs for fitting Poisson regression model

S. LALL , S. JAGGI, [1]E. VARGHESE, A. BHOWMIK AND C. VARGHESE

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110 012.
[1]ICAR-Central Marine Fisheries Research Institute, Kochi -682 018

Received : 11-01-2018 ; Revised : 15-03-2018 ; Accepted : 20-03-2018

ABSTRACT

Experiments related to herbicides or insecticides usually have the objective to find the effective concentration of the chemicals to control weeds or insects and to understand the relationship between the response and explanatory variables. The response is the number or proportion of organisms died and thus, is count data. The present study deals with the problem of developing experimental designs under Poisson regression model, which is a nonlinear model with count data as response. The focus here is to determine the unknown parameters of the model efficiently. The statistical designs generated are saturated and their performance is found better than traditionally used equally spaced designs. A simulation study is presented to demonstrate the application of the generated designs in actual experiment.

***Keywords*:** Saturated designs, D–optimality, Fisher information matrix

Weeds and pests are one of the most important factors to be considered for ensuring food security of the nation. Due to environmental hazards and pollution, environmental friendly measures are recommended for pest and weed management. However, agricultural chemicals used as pesticides or insecticides are still the most commonly used methods for controlling pests or weeds respectively and their use cannot be terminated permanently. The main motivation of the present study comes from two facts. The first fact is high dose results in increased mortality rate but is not recommended and the second fact is combination of two chemicals might give better results. Traditionally the focus was to find the dose of the chemical for which 50 per cent of the mortality rate is achieved and thus the response was taken as binary measurement like success or failure, 0 or 1, *etc*. Turner *et al*. (1992) studied herbicide Picloram at 4 doses to calculate its efficiency. A study on combination of two insecticides Pyrethrin and Piperonyl with the response as proportion of beetles died was reported by Hewlett (1969). In the present study, the response is taken as the count data where the number of organisms (pests or weeds) died is the response.

MATERIALS AND METHODS

This section formally explains the statistical model considered and the method of construction of experimental designs for fitting count data in agricultural experiments. A simulation procedure has also been devised to obtain empirical data under the considered model setup.

Poisson regression model

The Poisson regression model for two variables can be defined as:

$$y_i = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) = \exp[f(x,\theta)] = \eta(x,\theta) \quad (1)$$

where, y_i is the number of organisms died for the design point (x_{1i}, x_{2i}) and the unknown parameters are $\theta = [\beta_0, \beta_1, \beta_2]$ under the assumption that the effect of two predictors x_1 and x_2 is independent on the response.

Optimal designs

Let θ_0 be the initial parameter guess obtained from the previous experimental data or expert opinion and the objective is to find the design points or settings of the two predictors on which an experiment could be conducted to get the estimates of the unknown parameters in the model. A D-optimal design is the most suitable choice (Atkinson *et al*., 2007). Since precision in parameter estimates is required, the design with maximum information or minimum variance is preferred. D-optimal designs have the maximum determinant of Fisher Information Matrix (FIM) where FIM of a design (say ξ) under model (1) can be defined as:

$$\mathrm{M}(\xi) = \frac{1}{n}\sum_{i=1}^{n}\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \left.\begin{bmatrix} 1 & x_{1i} & x_{2i} \\ x_{1i} & x_{1i}^2 & x_{1i}x_{2i} \\ x_{2i} & x_{1i}x_{2i} & x_{2i}^2 \end{bmatrix}\right|_{\theta=\theta_0} \quad (2)$$

where, n is the number of design points.

Email: arpan.stat@gmail.com

Modified version of Fedorov algorithm can be employed to find the D-optimal designs under model setups (1) and (2) [see Fedorov (1972), Johnson and Nachtsheim (1983) and Dror and Steinberg (2006)]. Suppose the two chemicals or predictors x_1 and x_2 have the common range of 0 to 10 mg per cm^3. Under model setup (1), let θ_0 = [-16, 1, 1]. The choice of θ_0 is reasonable as the experimenter might not know the potency of the chemical and might guess $\beta_1 = \beta_2 = 1$. Since in equation (1), the expected response when both chemicals are applied at maximum concentration is assumed to be 60 and minimum possible response is 0, the value of β_0 is taken as -16. The locally D-optimal designs are D-optimal designs which depend upon the unknown parameters in the model. In a saturated design, the number of design points is equal to the number of parameters to be estimated in the model. Traditionally and intuitively experimenters often use equally spaced designs. Equally spaced designs have equidistant support points in the variable range and are obtained for the given number of support points (Dette *et al.*, 2008).

Construction of D-optimal designs

The basic logic behind the implemented algorithm is to find the best design with maximum determinant of FIM by simultaneously adding and removing a design point in a design. The algorithm needs an initial design and a set of potential design points also called as candidate set to start. Following are the key steps involved in this algorithmic approach:

1. Start with an initial design with positive determinant of corresponding FIM and a candidate set of potential design points.
2. At any stage or iteration, add a point to the existing design from the candidate set and simultaneously remove a point form the design itself so that the exchange gives maximum gain in determinant of FIM. The size of the design remains same throughout the procedure.
3. Step 2 is repeated for a fixed number of times or no further gain in determinant of IFM is observed.

Obviously the success of the algorithm depends on the suitable choice of candidate set and details regarding construction of efficient candidate sets can be found in Lall *et al.* (2018b). The considered algorithm has been described in detail in Lall *et al.* (2018a) for the case of logistic model. For the present study the relevant R codes developed for generating D-optimal designs have been given in appendix. Under this approach to find designs for agricultural experimental situations, the user should provide the range of explanatory variables and guesses for unknown parameters in the model. Using such designs with replications increase the precision in estimating parameters even under poor parameter guesses.

Simulation study

Since the reported designs are saturated, the designs are recommended to be used in replications so that at least 10 to 12 degrees of freedom is obtained for fitting the experimental data. The generated D-optimal design depends upon the choice of initial parameter guesses in this case [-16, 1, 1] and has only three design points. Suppose for actual experiment this design is replicated 5 times. For this final experimental setup, it is easy to simulate responses corresponding to the design points for some given values of initial parameter guesses. But simulating response with parameter values as initial guesses only is not appropriate. Unlike other situations with normally distributed response, an error or uncertainty cannot be added to a count data. So, a different approach is employed in the present study to mimic the actual experimental data and is explained below:

1. Choose the values of parameters randomly such that $\beta_0 \in [-20, -15]$, $\beta_1 \in [0.5, 1.5]$ and $\beta_2 \in [0.5, 1.5]$.
2. For a given design point say (x_1, x_2), generate a random Poisson number with mean = $\beta_0 + \beta_1 x_1 + \beta_2 x_2$.
3. Step 1 and 2 are repeated for about 2000 times to make a population of experimental data.
4. Now for the 15 design points obtained by taking 5 replications of the D-optimal design, chose a corresponding response form the population generated in Step 3.

RESULTS AND DISCUSSION

For the initial parameter guess θ_0 = [-16, 1, 1], the proposed algorithm generated following D-optimal design:

Table 1: Saturated experimental design

D–Optimal design	
x_1	x_2
10	10
8	10
10	8

The FIM for the design and its determinant is given below:

$$M(\xi) = \begin{bmatrix} 23.12 & 22.33 & 22.33 \\ 226.33 & 2223.87 & 2214.02 \\ 22.33 & 2214.02 & 2223.87 \end{bmatrix}, \quad |M(\xi)| = 1766.49$$

The design reported in table 1 has three design points and cannot be used directly for conducting the experiment. These design points are recommended to be replicated for 5 times randomly. The experimental data found through simulation is analyzed using 'glm' function of R software and the results are presented in table 2. The AIC value in this fitting is found to be 34.595. In the 'glm' procedure of R, there is provision for providing initial parameter guess. But this analysis in this study is done using default option in the R procedure. Although the design depends upon the values of initial parameter guesses, the analysis does not and all the coefficients are found to be significant based on the p-values. The estimates of the parameters and their respective standard errors are given in table 2. The mean response in model (1) is related to the parameters and explanatory variables through exponential function. The negative estimate of β_0 suggests a very low count of died organisms in the experiment under control condition when no chemical is applied. The positive estimates of β_1 and β_2 show that the both chemicals increase the count of died organisms with increase in application doses of the chemicals. The analysis of simulated data establishes the relationship between doses of chemicals and the count of died organisms in the considered experimental situation.

Table 2: Fitting of simulated experimental data

Coefficient	Estimate	Standard error	z-value	p-value
β_0	-20.5079	6.6516	-3.083	0.00205
β_1	0.8959	0.3819	2.346	0.01898
β_2	1.2425	0.5204	2.388	0.01696

In the present experimental situation, the equally spaced design found has three levels for both chemicals namely [0, 5, 10] and the design has 9 design points. The equally spaced design is compared with 3 replications of design reported in table 1 and the efficiency obtained is only 0.062 or 6.2 per cent.

A list of D-optimal designs for experiments with two chemicals based on Poisson regression model with respect to different settings of initial parameter guesses have been provided in table 3. Here, the range of both variables and chemicals are assumed to be [0, 10].

Table 3 gives some unique patterns in design search problem related to Poisson regression model with two predictors. It can be seen that the design points are not affected by the guess for parameter $\hat{a}_0$. All the designs have one common design point (10, 10) which implies that the maximum mortality is expected at maximum dose or concentration of both chemicals in the experiment. The determinants of FIMs are highly dependent on the choice of initial parameter guesses.

The results found in this study indicate the suitability of D-optimal designs in studies related to herbicides or insecticides when the objective is to fit the underlying model as precisely as possible. As precision and cost is very important in plant protection experiments, D-optimal designs can be recommended for decreasing the variance of parameter estimates in smaller number of runs. Even for the response measurements in the simulated experiments for parameters chosen randomly from a parametric space, it is seen empirically that the reported D-optimal design performs well and estimates the unknown parameters.

Table 3: D-optimal designs for poisson regression model with two variables

θ_0		D–Optimal Designs			|M|
(-15, 1, 1)	x_1	10	8	10	35480.97
	x_2	10	10	8	
(-16, 1, 1)	x_1	10	8	10	1766.49
	x_2	10	10	8	
(-17, 1, 1)	x_1	10	8	10	87.95
	x_2	10	10	8	
(-18, 1, 1)	x_1	10	8	10	4.38
	x_2	10	10	8	
(-19, 1, 1)	x_1	10	8	10	0.22
	x_2	10	10	8	
(-16, 0.5, 1.5)	x_1	10	10	6	3132.84
	x_2	10	8.6	10	
(-16, 0.5, 0.5)	x_1	10	10	6	2.644e-09
	x_2	10	6	10	
(-16, 0.5, 1)	x_1	10	10	6	0.002161498
	x_2	10	8	10	
(-16, 1, 0.5)	x_1	10	10	8	0.002161498
	x_2	10	6	10	
(-16, 1, 1.5)	x_1	10	10	8	2560329936
	x_2	10	8.6	10	
(-16, 1.5, 0.5)	x_1	10	10	8.6	3.32.84
	x_2	10	6	10	
(-16, 1.5, 1)	x_1	10	10	8.6	2560329936
	x_2	10	8	10	
(-16, 1.5, 1.5)	x_1	10	10	8.6	3.711e+15
	x_2	10	8.6	10	

ACKNOWLEDGEMENT

Authors are thankful to the editor and referee for their constructive suggestions for improving the quality of the manuscript. Authors are also thankful to Director, ICAR-IASRI, New Delhi for providing all kind of support to carry out the research work. First author would also like to thank PG School, IARI and ICAR-IASRI for providing senior research fellowship.

REFERENCES

Atkinson, A.C., Donev, A.N. and Tobias, R. 2007. *Optimum Experimental Designs with SAS.* Oxford University Press Oxford.

Dette, H., Kunert, J. and Peplyshev, A. 2008. Exact optimal designs for weighted least squares analysis with correlated errors. *Statistica Sinica*, **18**, 135-54.

Dror, H.A. and Steinberg, D.M. 2006. Robust experimental design for multivariate generalized linear models. *Technometrics*, **48**, 520-29.

Fedorov, V. V. 1972. *Theory of Optimal Experiments*, New York: Academic Press.

Hewlett, P. S. 1969. The toxicity to *Tribolium castaneum* (Herbst) (Coleoptera, Tenebrionidae) of mixtures of pyrethrins and piperonyl butoxide: Fitting a mathematical model. *J. Stored Prod. Res.*, ***5***(1), 1-9.

Johnson, M. E. and Nachtsheim, C. J. 1983. Some guidelines for constructing exact D-optimal designs on convex design spaces. *Technometrics*, **25**(3), 271-77.

Lall, S., Jaggi, S., Varghese, E., Varghese, C. and Bhowmik, A. D-Optimal designs for exponential and Poisson regression models. *J Indian Soc. Agric. Stat*.

Lall, S., Jaggi, S., Varghese, E., Varghese, C. and Bhowmik, A. 2018. An algorithmic approach to construct D-optimal saturated designs for logistic model. *J. Stat. Computation and Simulation*, **88**(6), 1191-99.

Russell, K.G., Woods, D.C., Lewis, S.M. and Ecclestom, J.A. 2009. D-optimal designs for Poisson regression models. *Statistica Sinica*, **19**, 721-730.

Turner, D. L., Ralphs, M. H., and Evans, J. O. 1992. Logistic analysis for monitoring and assessing herbicide efficacy. *Weed Technol*, **6**, 424-30.

Appendix

R codes for generation of D-optimal Designs

```
#########################
## making generic fx
fx<-function(q){
 q<-as.vector(q)
 q<-t(q)
 q0 <- q[,1]^0
 q11<-q[,1]^1
 q12 <- q[,1]^2
 q21<-q[,2]^1
 q22 <- q[,2]^2
 q3 <- q[,1]*q[,2]
 c<-rbind(q0,q11,q21)
 return(c)
}
f.exp<-function(q){
 beta<-beta.0
 x<-fx(q)
 z<-t(beta.0)%*%x
 z <- as.numeric(z)
 sqrt(exp(z)) * x
}
#########################
# generate design matrix
F.mat<-function(d0){
 m<-nrow(d0)
 p<- length(beta.0)  D<-matrix(0,ncol = m,nrow = p)
 for(i in 1:m)D[,i]<- f.exp(d0[i,])
 D
}
#########################
# compute std variance function for given point
d.fx<-function(q){
 f<-f.exp(q)
 d<-t(f)%*%solve(M)%*%f # M is the design
 return(d)
}

#########################
# Determinant of FIM
det.zi<-function(d0){
 N<-F.mat(d0)%*%t(F.mat(d0))
 M<-N/m
 det(M)
}
##########################
# Stopping Criterion
stop_crit<-function(d0){
 d1<-exc(d0)
 (det.zi(d1)-det.zi(d0))/det.zi(d0)
}
```