Tanzania Journal of Science 47(3): 999-1006, 2021
ISSN 0856-1761, e-ISSN 2507-7961
© College of Natural and Applied Sciences, University of Dar es Salaam, 2021

E-optimal Experimental Designs for Poisson Regression Models in Two and Three Variables

Emmanuel Idowu Olamide*, **Olusoga Akin Fasoranbaku and Femi Barnabas Adebola**
Department of Statistics, Federal University of Technology, Akure, Nigeria
**Corresponding author, e-mail: eiolamide@futa.edu.ng*
Co-authors' emails: oafasoranbaku@futa.edu.ng, fbadebola@futa.edu.ng

Received 24 Apr 2021, Revised 10 Jul 2021, Accepted 20 Jul 2021, Published Aug 2021
DOI: https://dx.doi.org/10.4314/tjs.v47i3.11

Abstract

In the context of generalized linear models, most of the recent studies were on logistic regression models and many of them focussed on optimal experimental designs with concentration on D-optimality. In this research, two- and three-variable Poisson regression models were considered for E-optimization on restricted design space [0, 1]. The two-variable Poisson regression model was not optimal at 3-design points, but was found to be E-optimal at 4-design points (1, 1), (0, 0), (0, 1) and (1, 0) with equal design weights of 0.25. The three-variable Poisson regression model was E-optimal at 4-design points (0, 0, 1), (0, 1, 0), (1, 1, 1) and (1, 0, 0) with each design point having design weights of 0.25. The prediction error variance (PEV) for the two-variable Poisson regression model is 0.35 and that of the three-variable Poisson regression model is 0.68. From this research, the two-variable Poisson regression model is preferred to the three-variable Poisson regression model because of smaller PEV.

Keywords: E-optimality, Fisher Information Matrix, Poisson Regression Model, Prediction Error Variance.

Introduction

Arrays of designs with high efficiency in relation to some statistical measures are referred to as optimal designs. The basics of experimental designs originated from Smith (1918) in his innovative mathematical derivation. The significance of nonlinear models in applied fields cannot be underestimated. Experimental designs for practical situations usually comprise at least a number of nonlinear components. The ease that comes with the construction of optimal designs for linear models is not valid for nonlinear models due to parameter dependency in the latter. Optimal experimental designs play vital roles in several fields of applications. For instance, they are widely applied in medicine, biology, agriculture and industries. Generation of optimal experimental designs is model-dependent and optimization process involves the Fisher information matrix. For example, when examining a compound in dose-response studies, adequate knowledge in addition to proper characterization of dose as well as its reactions is a major step to be considered because poor understanding of the dose response records can have an undeviating effect when estimating the chosen level of dose. In the case of drug development settings, choosing a very high quantity of dose may lead to intolerable toxicity and harmfulness, while selecting very few quantity of dose can reduce the possibility of having effectiveness in the confirmatory stage. This can therefore reduce the possibility of obtaining endorsement and approval for the drug from the regulatory body.

Studies on optimal designs commenced with Smith (1918) through his amazing paper on the generation of G-optimum designs for a sixth-order polynomial model in one variable. T-optimum designs for integrated variance are designs concerned with the minimization of variance component in the theoretical framework of optimum experimental design (Studden 1977). Chaudhuri and Mykland (1993) examined proper design of nonlinear experiments that will enable the construction of efficient parameter estimates. Emphasis was on a very broad nonlinear structure which includes many models that are usually faced in practice. Two essential stages were considered for the experiments: the static design phase and completely adaptive sequential phase where support points were sequentially selected to explore design optimality for D-criterion through the estimates of parameters obtained from accessible information. Exploration of the performance of maximum likelihood estimator using the generated data from such experiment was considered. The two core procedural obstacles encountered are the nature of data dependency from the adaptive sequential experiment as well as the experimental randomness contained in the overall Fisher information. Likelihood-based structure of martingale was explored through the analysis. Derivations of necessary conditions in ensuring convergence of the selected design to D-optimality as first trial increases were considered. The average Fisher information converges to provide a condition of ergodicity associated with the martingale process growth and has intrinsic association with the likelihood as well as ensuring optimality of the design for large sample. This major observation ultimately produced the first-order efficiency of estimate obtained by the maximum likelihood technique through the central limit theorem of martingale process and the validation of statistical inference for large sample based on the likelihood was confirmed. Krewski et al. (2002) developed optimal designs for estimating effective dose in growing toxicity. The dose-response for prenatal death and foetal malformation were jointly modelled through the Weibull distribution. Approximate optimal designs were generated for prenatal death, malformation and total toxicity, especially when the series of developmental studies were extended. The designs comprise three groups of doses, which include: the control group, the low and high doses. The effects on optimal designs when the number of implants and degree of intra-litter correlation are varied were examined. Though, only three dose groups are being considered for optimality in most cases, considerations of practicability, especially when it involves the estimation of the dose-response curve shape and lack of fit of the model endorses the use of suboptimal designs involving more than three doses in practice. Han and Chaloner (2003) considered exponential decay models involving one, two and three parameters for the derivation of locally D- and C-optimal designs using analytical approach. The locally optimal designs were observed to be invariant to reparameterization. The approach was illustrated via Bayesian optimal designs. Myung and Pitt (2009) observed that experimental discrimination among models of psychological processes is problematic because of the difficulty in determining the values of the critical factors that provides most information in differentiation. Possible determination of the values can be accounted for through current advancements in sampling-based search approaches, thereby leading to identification of an optimal experimental design. Demonstration was considered through application of the method on retention and categorization that constitute two gratified areas of studies in cognitive psychology wherein models are competitively feasible. The quality of designs considered in literature was compared with the optimal design. From the results, the efficiency of experimental method is potentially increased through design optimization. Yang and Stufken (2009) proposed a new technique in identifying support points of a locally optimum design that pertains to a nonlinear model. The basis of the approach forms the algebraic features and was comparatively studied with the frequently used geometric method. Models containing two parameters were considered and applications of the general findings to common special cases of some nonlinear regression models. The Michaelis-Menten, probit, logistic, double reciprocal, double exponential and a loglinear Poisson regression models were the nonlinear models considered. The technique which is greatly important in conducting multi-stage experiments performs well with both restricted as well as unrestricted design spaces and can be easily and relatively implemented. Dette et al. (2010) estimated the slope of mean response in a regression model through experimental designs. The locally and standardized minimax optimal designs were fully discussed. A generalization of the findings concerning the number of support points of locally optimal designs was produced through the formation of a Chebyshev system from the regression functions. Polynomial and Fourier regression models of arbitrary degrees were explicitly considered for the construction of optimal designs in estimating the slope of the regression function.

Burghaus and Dette (2014) investigated design optimality using Bayesian approach with non-informative prior distributions. The Berger-Bernardo and Jeffreys priors for non-concave optimality criteria were particularly studied. Boukouvalas et al. (2014) considered a normal linear regression model having input-dependent noise for the generation of optimal design for parameter estimation. The field of computer experiments, where simulators that are computationally challenging are approximated by means of normal emulators in acting as statistical surrogates motivated the research. Recurrent assessments play supportive role by using replicated observations in experimental designs especially for stochastic simulators that yield varying responses for some model inputs. The normal regression and kriging models were widely considered in the framework of experimental design for application. Minimization of the variance of estimated normal parameters forms the basis for generating designs. A normal linear model with heteroscedasticity was considered for optimization. The approximation error of the variance of parameters is reduced through the inverse of the Fisher information as replication points increased, which was shown through empirical studies. Results from series of simulation experiments on both synthetic and systems biology data showed that optimal designs with replicated observations performed better than space-filling designs. A major review on E-optimality is the work of Dette et al. (2004) when a general set of nonlinear regression model for the investigation of the local E- and C-optimal designs were considered. The Chebyshev points representing the local extrema of the equi-oscillating best approximation of the function f0≡0 through a normalized linear combination of the regression functions in the equivalent linearized model provide the support points generated for the E- and C-optimal design criteria in several cases. Logistic, exponential and rational models were the considered models. The E- and C-optimal design problems were solved explicitly in several cases for the rational regression models.

Most researches on optimal design of experiments focus on the D-optimality criterion. This paper examines the E-optimal design criterion which aids the maximization of the least eigenvalue of the information matrix.

Materials and Methods

Poisson regression models

Poisson regression model can be largely written as:

$$y_{ij} \sim Poisson\ (\tau_i) \qquad (1)$$

The mean response τ_i can be expressed as:

$$\tau_i \ = exp(X_i' \beta) \qquad (2)$$

where, y_{ij} are the response variables, τ_i is the expectation of the response variable at the i^{th} design point, X_i' is the design matrix containing factors X_i (i = 1, 2, …), and β is a vector of parameters.

Construction of E-optimal designs for two-variable Poisson regression model

A typical Poisson regression model in two variables can be represented by:

$$\tau_i = exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \qquad (3)$$

A basic assumption of a Poisson regression model is the positive nature of the response variables.

An optimal or a near-optimal design, $\xi \in \Xi$, in design space, χ, containing definite design points is denoted by:

$$\xi = \begin{Bmatrix} x_1, x_2, \cdots, x_s \\ w_1, w_2, \cdots, w_s \end{Bmatrix} \qquad (4)$$

where, $x_i \in \chi$ (the support points) is a compact subset of real numbers, w_i are the weights of the design at each support point satisfying $0 < w_i \leq 1$ and $\sum_{i=1}^{s} w_i = 1$.

Considering the two-variable Poisson regression model in Equation (3),

$$ln\, \tau_i = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \qquad (5)$$

Here,

$$f'(x_i) = (1, \quad x_1, \quad x_2) \qquad (6)$$

where, $f'(x_i)$ is the i^{th} row of X, a known function of predictor variables.

The element of the Fisher information matrix is therefore obtained and presented as:

$$\mathrm{M} = \mathrm{X'X} = \begin{bmatrix} 1 & x_1 & x_2 \\ x_1 & {x_1}^2 & x_1x_2 \\ x_2 & x_1x_2 & {x_2}^2 \end{bmatrix} \qquad (7)$$

The Fisher information matrix can be expressed in compact form as:

$$\mathrm{M}(\xi;\ \beta_0, \beta_1, \beta_2) = \sum w_i \tau_i f(x_i)\, f'(x_i) \qquad (8)$$

and more compactly as:

$$\mathrm{M}(\xi;\ \beta_0, \beta_1, \beta_2) = X'WX \qquad (9)$$

where, w_i represents the weights of the support points, $\tau_i = \exp(\eta_i)$, is the mean response of the ith design point, ξ is the design measure, and $W = diag\{w_i\, \mu_i\}$, and $X = [f(x_1),\ f(x_2)]$.

Explicitly, the Fisher information matrix for Equation (3) is therefore obtained as:

$$\mathrm{M}(\xi;\ \beta_0, \beta_1, \beta_2) = \begin{bmatrix} \sum w_i\tau_i & \sum w_i\tau_i\, x_{1\mathrm{i}} & \sum w_i\tau_i\, x_{2\mathrm{i}} \\ \sum w_i\tau_i\, x_{1\mathrm{i}} & \sum w_i\tau_i\, {x_{1\mathrm{i}}}^2 & \sum w_i\tau_i\, x_{1\mathrm{i}}x_{2\mathrm{i}} \\ \sum w_i\tau_i\, x_{2\mathrm{i}} & \sum w_i\tau_i\, x_{1\mathrm{i}}x_{2\mathrm{i}} & \sum w_i\tau_i\, {x_{2\mathrm{i}}}^2 \end{bmatrix} \qquad (10)$$

Suppose the Eigenvalue of the Fisher information matrix in Equation (10) is λ_i, the E-optimality design criterion seeks the minimization of the variance of the least well estimated linear combination $a^T\hat{\beta}$ conditionally upon the constraint that $a^Ta = 1$. It maximizes the minimum eigenvalue of the information matrix. This equivalently minimizes the maximum relating to the inverse of eigenvalue of the information matrix.

i.e.,

$$E - optimal = \min \max_i \frac{1}{\lambda_i}.$$

Construction of E-optimal designs for three-variable Poisson regression model

A three-variable Poisson regression model can be defined as:

$$\tau_i = exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}) \qquad (11)$$

The same procedure used in obtaining the Fisher information matrix and eigenvalue for the two-variable Poisson regression model in Equation (3) is employed in the case of a three-variable Poisson regression model presented in Equation (11).

Thus, the Fisher information matrix is obtained as:

$$\mathrm{M}(\xi, \beta_0, \beta_1, \beta_2, \beta_3) = \begin{bmatrix} \sum_{i=1}^{4} w_i\tau_i & \sum_{i=1}^{4} w_i\tau_i\, x_{1\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, x_{2\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, x_{3\mathrm{i}} \\ \sum_{i=1}^{4} w_i\tau_i\, x_{1\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, {x_{1\mathrm{i}}}^2 & \sum_{i=1}^{4} w_i\tau_i\, x_{1\mathrm{i}}x_{2\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, x_{1\mathrm{i}}x_{3\mathrm{i}} \\ \sum_{i=1}^{4} w_i\tau_i\, x_{2\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, x_{1\mathrm{i}}x_{2\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, {x_{2\mathrm{i}}}^2 & \sum_{i=1}^{4} w_i\tau_i\, x_{2\mathrm{i}}x_{3\mathrm{i}} \\ \sum_{i=1}^{4} w_i\tau_i\, x_{3\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, x_{1\mathrm{i}}x_{3\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, x_{2\mathrm{i}}x_{3\mathrm{i}} & \sum_{i=1}^{4} w_i\tau_i\, {x_{3\mathrm{i}}}^2 \end{bmatrix} \quad (12)$$

In terms of eigenvalues,
$E - optimal = \ \min \max_i \frac{1}{\lambda_i}$.

Results and Discussion

E-optimal designs for two-variable Poisson regression model

The constructed E-optimal designs relating to Poisson regression model with two predictor variables in Equation (3) is presented as:

$$\xi_E^* = \begin{Bmatrix} (1, \quad 1) & (0, \quad 0) & (0, \quad 1) & (1, \quad 0) \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix} \quad (13)$$

Considering the two-variable Poisson regression model, the design is not optimal at 3-point design, which necessitated an increase in the number of design points. At 4-point design, the design is found to be E-optimal. After 1000 iterations, the optimal design points are $x_1 = 1, x_2 = 1$; $x_1 = 0, x_2 = 0$; $x_1 = 0, x_2 = 1$; and $x_1 = 1, x_2 = 0$. The constructed E-optimal design weights at each optimal design point are $w_1 = 0.25$, $w_2 = 0.25$, $w_3 = 0.25$, and $w_4 = 0.25$, respectively.
This means that 25% of the total experimental runs are allocated to each optimal design point.

Figure 1 shows the E-optimal criterion value to be 4.0. The positive nature of this value and being $\geq p$, supports the choice of the design space considered in this study.

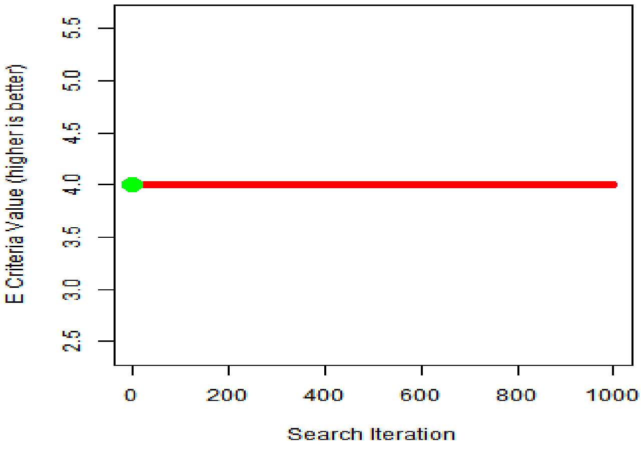

Figure 1: E-optimal criterion value for two- variable Poisson regression model.

Figure 2 gives the prediction error variance of the two-variable Poisson regression model to be 0.35, which verifies the E-optimality of the design at 4-design points.

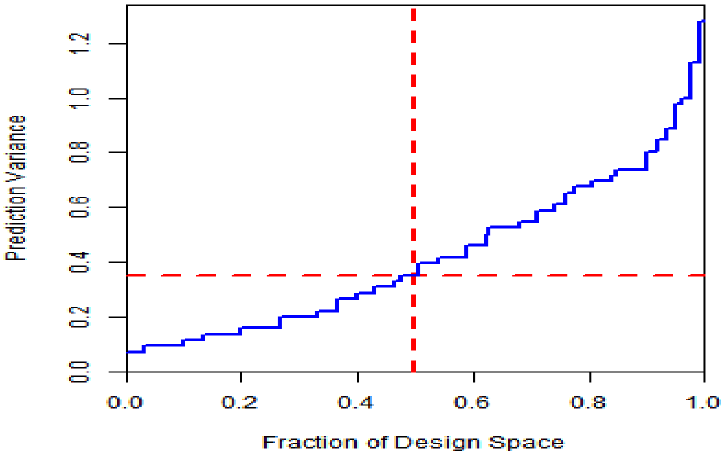

Figure 2: Prediction error variance of e-optimal design for two-variable Poisson regression model.

E-optimal designs for three-variable Poisson regression model

The findings of E-optimal designs for the three-variable Poisson regression model in Equation (11) are presented as:

$$\xi_E^* = \begin{Bmatrix} (0,\ 0,\ 1) & (0,\ 1,\ 0) & (1,\ 1,\ 1) & (1,\ 0,\ 0) \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix} \tag{14}$$

The design is E-optimal at 4-design points. After 1000 iterations, the optimal design involves collection of optimal points $x_1 = 0, x_2 = 0, x_3 = 1$; $x_1 = 0, x_2 = 1, x_3 = 0$; $x_1 = 1, x_2 = 1, x_3 = 1$; and $x_1 = 1, x_2 = 0, x_3 = 0$. The generated E-optimal design weights at each optimal design point are $w_1 = 0.25$, $w_2 = 0.25$, $w_3 = 0.25$, and $w_4 = 0.25$, respectively. This means that 25% of the total experimental runs are allocated to each optimal design point.

Figure 3 shows the E-optimal criterion value to be 4.0. The positive nature of this value and being $\geq p$, corroborates the choice of the design space considered in this study.

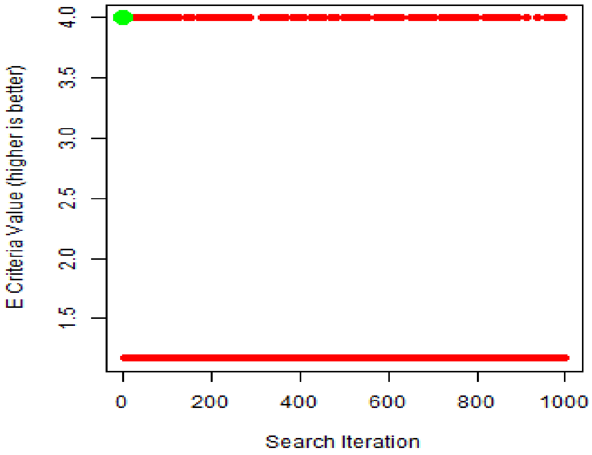

Figure 3: E-optimal criterion value for three-variable Poisson regression model.

Figure 4 gives the prediction error variance of Equation (11) to be 0.68, which confirms that the design is indeed E-optimal at 4-design points.

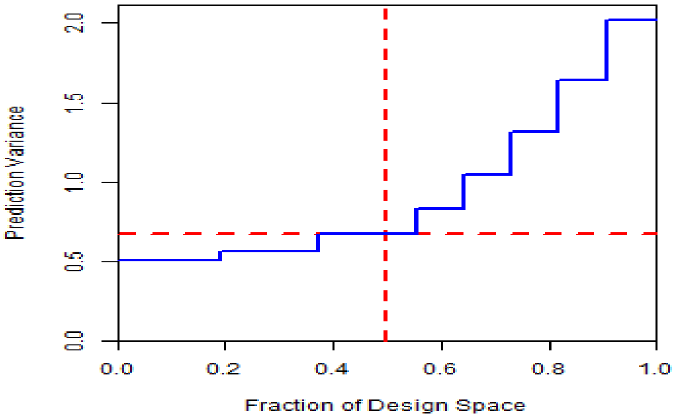

Figure 4: Prediction error variance of E-optimal design for three-variable Poisson regression model.

Conclusion

This research investigated and generated E-optimal experimental designs for Poisson regression models containing two and three variables in linear terms. Both the two- and three-variable Poisson regression models were found to be E-optimal at 4-design points with equal weights and they both have equal criteria values. The prediction error variance for the two-variable Poisson regression model was observed to be smaller than that of the three-variable Poisson regression model, thus making the two-variable Poisson regression model more preferred.

References

Boukouvalas A, Cornford D and Stehlík M 2014 Optimal design for correlated processes with input-dependent noise. *Comput. Stat. Data Anal.* 71: 1088-1102.

Burghaus I and Dette H 2014 Optimal designs for nonlinear regression models with respect to non-informative priors. *J. Stat. Plan. Infer.* 154: 12-25.

Chaudhuri P and Mykland PA 1993 Nonlinear experiments: optimal design and inference based on likelihood. *J. Am. Stat. Assoc.* 88(422): 538–546.

Dette H, Melas VB and Pepelyshev A 2004 Optimal designs for a class of nonlinear regression models. *Ann. Stat.* 32 (5): 2142-2167.

Dette H, Melas VB and Pepelyshev A 2010 Optimal designs for estimating the slope of a regression. *Statistics* 44(6): 617-628.

Han C and Chaloner K 2003 D- and c-optimal designs for exponential regression models used in viral dynamics and other applications. *J. Stat. Plan. Infer.* 115(2): 585-601.

Krewski D, Smythe R and Fung K 2002 Optimal designs for estimating the effective dose in developmental toxicity experiments. *Risk Anal.: Int. J.* 22 (6): 1195-1205.

Myung JI and Pitt MA 2009 Optimal experimental design for model discrimination. *Psychol. Rev.* 116 (3): 499-518.

Smith K 1918 On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* 12: 1-85.

Studden WJ 1977 Optimal designs for integrated variance in polynomial regression. In: Shanti S, Gupta, David S. Moore (Eds), *Proceedings of a Symposium Held at Purdue University* May 17–19, 1976: *Statistical Decision Theory and Related Topics* (pp. 411-420), Academic Press.

Yang M and Stufken J 2009 Support points of locally optimal designs for nonlinear models with two parameters. *Ann. Stat.* 37(1): 518-541.