Tanzania Journal of Science 47(3): 1303-1309, 2021
ISSN 0856-1761, e-ISSN 2507-7961

Weighted Statistical A-optimal Designs for Two-variable Poisson Regression Model with Application to Fertility Studies

Emmanuel Idowu Olamide*, Femi Barnabas Adebola and Olusoga Akin Fasoranbaku
Department of Statistics, Federal University of Technology, Akure, P.M.B. 704, Akure, Ondo State, Nigeria
** Corresponding Author. E-mails: eiolamide@futa.edu.ng, eiolamide@gmail.com; fbadebola@futa.edu.ng; oafasoranbaku@futa.edu.ng*
Received 19 Feb 2021, Revised 19 Aug 2021, Accepted 19 Aug 2021, Published Aug 2021
DOI: https://dx.doi.org/10.4314/tjs.v47i3.37

Abstract

This research extends design optimization to model involving count data. A two-variable Poisson regression model was investigated for A-optimality on a constrained design space and the weights of the optimal design points were obtained. The constructed designs were verified to be A-optimal at 4-point design through the general equivalence theorem. The efficiency of the constructed optimal design was found to be 100% A-efficient. The concept of weighted optimal designs for Poisson regression model was applied to fertility studies. Approximate A-optimal design weights of educational level of women were obtained for each marriage duration period with respect to their places of residence. The study revealed that the numbers of women with secondary education and above were found to be consistently more than that of women with no education, lower primary education and upper primary education, respectively for all the marriage duration periods considered and at each place of residence. The only exclusion is the marriage duration of 0–4 years at Suva where the proportion of women with no education was more than other educational levels.

Keywords: A-optimality, Design Point, Fisher Information Matrix, Imperialist Competitive Algorithm, Poisson Regression Model.

Introduction

Optimal designs of experiments are a set of designs that produce maximum efficiency with respect to some statistical criteria. Here, efficiency refers to the degree of the worth of experimental design, that is, the precision of variances of estimators when dealing with estimation of model parameters. Classification of optimal designs is based on the concept that provides a methodical approach of obtaining the possible best or a highly efficient design using available data with respect to the situation being considered. Optimal experimental design is an essential area of scientific research. The fundamental idea underlying optimality theory of a design is that statistical inferences about quantities of interest can be improved by optimally selecting levels of the control variables at low costs (Berger and Wong 2009).

Kiefer (1959) originally conceptualized the development of D-optimal designs majorly for models of response surfaces with the review of current advancements laying emphasis on prospective and actual effectiveness. Emphases were majorly on linear models for the generation of exact designs, especially when considering design regions with irregularities, just as seen in the blocking of response surface designs. Systematic designs and other robust areas as well as mixture designs with irregular design regions were also emphasized. A non-

linear model was considered for the generation of D- and C-optimal designs using the cereal's economic response of production to fertilizers' levels, with conditions of maximum economic return generated for the C-optimum design. Descriptions of locally and Bayesian designs were considered. Designs for low-dose-95 (LD95) in a logistic version of a generalized linear model were the aftermath of the related findings where observations on different responses from genders were considered. Designs with structured variances were proposed as substitutes for the potentially inefficient Taguchi product design and construction of designs pertaining to clinical trials with random balances were considered (Atkinson 1996). Construction of G-optimum designs for a sextic polynomial model in one variable forms the pioneer work on optimal designs, though immediate effect was not recorded from the study but the designs were indeed optimal (Smith 1918). A comparative study of some designs with the aid of the determinant of Fisher information matrix produces D-optimal design (Wald 1943).

Development of Mathematical theory of weighing designs was examined by Hotelling (1944) and Mood (1946). The concept of design optimality is applicable to models involving design variables of first-order term and biases due to the omission of higher-order terms are avoidable through design rotation (Box 1952). A two-variable regression model that is devoid of intercept was examined for C- and A-optimal design criteria (Elfving 1952). The findings were generalized for non-linear models with the dependence of the design on unknown parameter values of the non-linear model justifying the local optimality of the design (Chernoff 1953). D-optimum design is independent of model parameters in an experiment involving effective dose levels in generating optimum design for one-variable first-order Poisson model. Investigations on the dependence of D-optimal designs concerning model parameters of the research interest with a quadratic term in one variable, as well as with additive two-variables with interaction term was considered and the performances of certain appealing standard designs examined (Russell et al. 2009). Haines et al. (2007) constructed D-optimal designs for logistic regression model, numerical illustration of the design optimality was examined over a range of intercept values and the singular case of the cut-off value of intercept was algebraically proved. Yang et al. (2011) proposed a new method of identifying optimal designs relating to the probit and logistic models with many factors. D-, A- and E- optimal designs were explicitly formulated with respect to optimality and the general structure of the optimal designs was identified.

A mechanistic, empirical or hybrid of the two constitute a nonlinear model that can be conveniently fitted for several chemical and biological experiments which comprise of multiple treatment factors. Conventional point and coordinate exchange algorithms can first be implemented in order to calculate as well as examine optimal designs. An innovative multiphase optimisation technique in constructing D-optimal designs with improved properties was developed. The advantages of the technique were demonstrated in applying it to two experiments that involve nonlinear regression models. The generated designs were observed to be substantially more informative than designs generated through traditional design optimality algorithms (Huang et al. 2019).

A properly planned experiment provides solutions to queries in an obvious, succinct and effective manner. Additional information to simple plots is usually needed for the initial analysis in revealing the nature of dependence of responses on design factors. Consequently, the use of least squares is often required in estimating the parameters in order to demonstrate the dependence. Good experiments are expected to produce results of estimated parameters with minimum variances and covariances. Functions of variances are minimized by optimally designing experiments, thereby aiding provision of reasonable parameter estimates and predictions of responses (Atkinson 2015).

Dependent and independent variables constitute nonlinear experiments and are related via a framework of nonlinear regression-type model. The following regressions: Poison, Probit, Gamma, inverse Gaussian, and the like, constitute typical nonlinear models. A complex nonlinear function of the unknown parameter of interest constitutes the associated Fisher information of nonlinear experiment. There is an inflexible situation when designing optimal experiments for nonlinear model since efficient experimental design requires prior parameter knowledge, whereas, an experiment is essentially designed for the purpose of data generation intended for estimation of parameters. Optimal designs for logistic regression model has received increasing attention in recent years (Wang et al. 2006), our attention is hereby drawn to the construction of optimal designs for model involving count variables.

This study considers the construction of Weighted A-optimal designs for a two-variable Poisson regression model with application to fertility studies.

Materials and Methods

Poisson regression model

The Poisson regression model can be broadly written as in Equation (1).

$$y_{ij} \sim Poisson\ (\mu_i) \quad (1)$$

The mean response, μ_i, can be expressed as in Equation (2);

$$\mu_i = exp(X_i' \beta) \quad (2)$$

where, y_{ij} are the response variables, μ_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.

This research focuses on the development of optimal experimental designs subject to the multiple linear Poisson regression model in Equation (3).

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

The assumption for a Poisson regression model without loss of generality is that the response variables are nonnegative.

Approximate design, $\xi \in \Xi$, in design space, χ, containing definite design points is denoted by Equation (4);

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

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

Imperialist competitive algorithm

The imperialist competitive algorithm (ICA) is a computational technique employed in solving different kinds of optimization problems. It is a mathematical model and computer assimilation of human social evolution. The ICA can be defined as a form of meta-heuristic algorithm designed for solving optimization problems. In design of experiments, the imperialist competitive algorithm is basically used to solve optimal design problems pertaining to non-linear models (Kaveh and Talatahari 2010). Since the criteria of optimality for nonlinear models depends on the unknown parameters, the locally optimal design function deals with the parameter-dependency based on the information available for the unknown parameters.

Fisher information matrix

Fisher information matrix is expressly defined in terms of a design measure expressed as $M(\xi; \beta)$.

Considering the model in Equation (3),

$$ln\ \mu_i = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \quad (5)$$

Here,

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

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 as in Equation (7).

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

The Fisher information matrix can be expressed in compact form as in Equation (8);

$$M(\xi; \beta_0, \beta_1, \beta_2) = \sum w_i \mu_i f(x_i) f'(x_i) \quad (8)$$

and more compactly, as in Equation (9);

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

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

Explicitly, the Fisher information matrix for the model in Equation (3) can therefore be expressed as in Equation (10).

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

A-optimality ("average" or trace)

A-optimal criterion searches for the minimization of the trace of the inverse of information matrix, i.e.,

$\underset{x_i, i=1,\cdots,n}{min}\ trace\ (X'X^{-1})$.

The A-optimal criterion is equivalent to the minimization of the average variance of estimates of the regression parameters.

Construction of A-optimal designs

The A-optimality design criterion seeks the minimization of the trace relating to the inverse of the information matrix. In other words, the sum of the variances of the parameter estimates is minimized, equivalently minimizing the average variance. Considering the two-variable Poisson regression model in Equation (3), suppose information matrix in Equation (10) has eigenvalues, λ_i, then, the expression for A-optimality in terms of the eigenvalues is as expressed in Equation (11);

$$A - optimal = min \sum_{i=1}^{p} \frac{1}{\lambda_i} \qquad (11)$$

where, λ_i are the eigenvalues of the information matrix, and p is the number of parameters.

Results

A-optimal designs for two-variable Poisson regression model

The results of A-optimal design pertaining to a Poisson regression model with two predictor variables is presented in Equation (12).

$$\xi_A^* = \begin{Bmatrix} (0,\ 0) & (0,\ 1) & (1,\ 0) & (1,\ 1) \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix} \qquad (12)$$

The two-variable additive Poisson regression model in Equation (3) is A-optimal at 4-design points as shown in Equation (12). After 1000 iterations, the A-optimal criterion value is 100 and the optimal design points are $x_1 = 0,\ x_2 = 0;\quad x_1 = 0,\ x_2 = 1;\quad x_1 = 1,\ x_2 = 0$ and $x_1 = 1,\ x_2 = 1$. The constructed A-optimal design weight at each optimum design point is $w_1 = w_2 = w_3 = w_4 = 0.25;$ which implies that 25% of the total experimental runs is allocated to each optimum design region.

Figure 1 presents the prediction variance of A-optimal design relating to Poisson regression model involving two predictor variables. The prediction variance is smaller, as compared to the number of parameters in the model, which verifies the design-optimality at 4-design points.

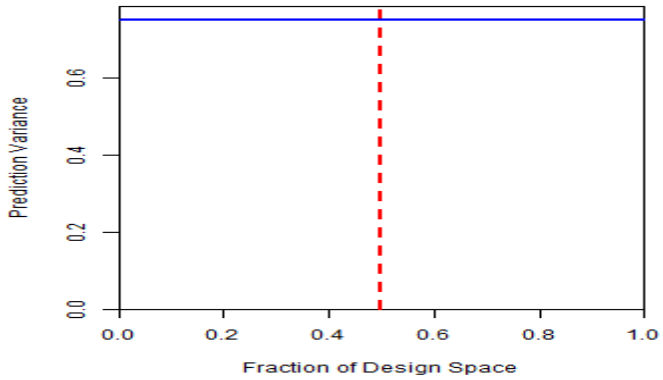

Figure 1: A-optimal prediction variance for two-variable Poisson regression model.

Figure 2 presents the optimal efficiency search values of the design relating to Poisson regression model containing two design variables. The design is observed to be perfectly A-efficient after 1000 search iterations.

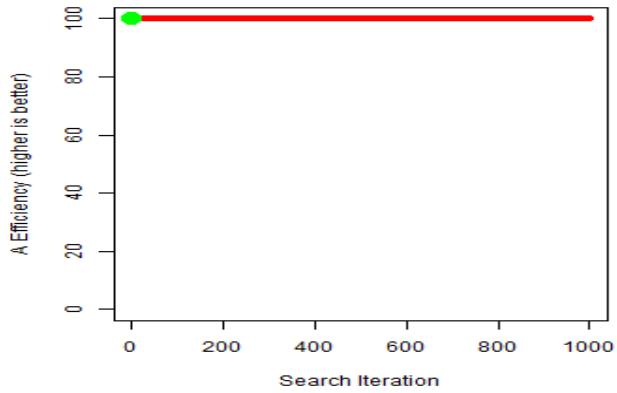

Figure 2: Optimal efficiency search values of A-optimal design for Poisson regression model in two variables.

Applications in fertility studies

The procedure that generates the A-optimal design through best parameter guess approach is hereby applied in fertility survey. The children ever born data extracted from Rodriguez (2007) for Indian women from world fertility survey are considered in this study. Herein, the response variable is the number of children born by the Indian women, while the predictor variables are the marriage duration, places of residence and educational levels of the women. The focus here is the construction of the A-optimal design weights of educational levels of the women for each marriage duration period with respect to their places of residence. The results are presented in Tables 1 and 2, respectively.

Table 1: Weights of A-optimal designs in fertility survey

	Place of residence											
	Suva				Urban				Rural			
Mar. Dur.	N	LP	UP	S+	N	LP	UP	S+	N	LP	UP	S+
0 - 4	0.3091	0.2047	0.2304	0.2558	0.2207	0.2589	0.2330	0.2874	0.2410	0.2422	0.2410	0.2759
5 - 9	0.2151	0.2318	0.2652	0.2879	0.1993	0.2608	0.2594	0.2806	0.2508	0.2380	0.2493	0.2618
10 - 14	0.2142	0.2258	0.2541	0.3059	0.2319	0.2596	0.2489	0.2596	0.2415	0.2415	0.2476	0.2693
15 - 19	0.2317	0.2139	0.2678	0.2866	0.2463	0.2307	0.2490	0.2740	0.2173	0.2067	0.2304	0.3456
20 - 24	0.2118	0.2232	0.2536	0.3114	0.2504	0.2391	0.2593	0.2511	0.2138	0.2158	0.2268	0.3436
25 - 29	0.2035	0.2014	0.2254	0.3697	0.3494	0.3256	0.3250	-	0.3211	0.3142	0.3647	-

Where, N is no education, LP is lower primary education, UP is upper primary education, and S+ is secondary education and above.

Table 2: Optimal experimental designs of educational levels of women in fertility survey

	Place of residence											
	Suva				Urban				Rural			
Mar. Dur.	(0, 0)	(0, 1)	(1, 0)	(1, 1)	(0, 0)	(0, 1)	(1, 0)	(1, 1)	(0, 0)	(0, 1)	(1, 0)	(1, 1)
0 - 4	0.3091	0.2047	0.2304	0.2558	0.2207	0.2589	0.2330	0.2874	0.2410	0.2422	0.2410	0.2759
5 - 9	0.2151	0.2318	0.2652	0.2879	0.1993	0.2608	0.2594	0.2806	0.2508	0.2380	0.2493	0.2618
10 - 14	0.2142	0.2258	0.2541	0.3059	0.2319	0.2596	0.2489	0.2596	0.2415	0.2415	0.2476	0.2693
15 - 19	0.2317	0.2139	0.2678	0.2866	0.2463	0.2307	0.2490	0.2740	0.2173	0.2067	0.2304	0.3456
20 - 24	0.2118	0.2232	0.2536	0.3114	0.2504	0.2391	0.2593	0.2511	0.2138	0.2158	0.2268	0.3436
25 - 29	0.2035	0.2014	0.2254	0.3697	0.3494	0.3256	0.3250	-	0.3211	0.3142	0.3647	-

Discussion

Tables 1 and 2 present the findings for the application of optimal designs employed in the construction of the proportion of educational levels of women with respect to their places of residence for each marriage duration period. The weights of their educational levels were generated for each marriage duration period. Considering the sample size for each marriage duration period with corresponding place of residence, the number of women of child-bearing age, having acquired at least secondary education tends to be the highest in each place of residence for all marriage duration periods except in Suva, where women with at most four years of marriage duration tended to record the highest number of residents with no education. This may be practically due to the influx of migrants from rural areas to urban and the nation's capital city.

Generally, majority of the parents tend to morally and educationally ensure that their female wards acquire at least secondary education. Social and technological advancements can be said to have accounted for the reasonable low numbers of women with no education, lower primary education and upper primary education.

Conclusion

This research investigated and generated A-optimal experimental designs for Poisson regression model containing two predictor variables in linear terms. The A-optimal design

criterion was constructed for the model and the condition for optimality was verified via the general equivalence theorem. The optimality criterion considered in this work was found to be indeed optimal. The efficiency of the optimality criterion was also established and found to be very efficient. The concept of optimal experimental design was applied to fertility studies using the children ever born data surveyed on Indian women race. Approximate A-optimal design weights were constructed for each marriage duration period at each place of residence using their levels of education. From this research, it can be deduced that the two-variable Poisson regression model was found to be 100% A-efficient at 4-point design. In fertility studies, design optimality has been shown in this research to play important roles in determining the proportion of educational level of women with respect to their marriage durations and place of residence. This study therefore recommends that the government or concerned agency in policy making should improve on the quality of education pertaining to child bearing, as this will increase and improve the educational levels of women of child bearing age particularly in rural areas.

Acknowledgement

We are grateful to the Tertiary Education Trust Fund unit of the Federal Republic of Nigeria for funding this research with the aid of the 2017 institution based research grant.

Conflict of Interest

The authors declare no conflict of interest.

References

Atkinson AC 1996 The usefulness of optimum experimental designs. *J. Royal Stat. Soc. Ser. B* 58: 59-76.

Atkinson AC 2015 International encyclopedia of the social and behavioural sciences. Elsevier Science Ltd.

Berger MPF and Wong W 2009 An introduction to optimal designs for social and biomedical research. *Wiley*.

Box GEP 1952 Multi-factor designs of first order. *Biometrika* 39: 49-57.

Chernoff H 1953 Locally optimal designs for estimating parameters. *Ann. Math. Stat.* 24: 586-602.

Elfving G 1952 Optimum allocation in linear regression theory. *Ann. Math. Statist.* 23(2): 255-262.

Haines LM, Kabera G, Ndlovu P and O'Brien TE 2007 D-optimal designs for logistic regression in two variables. *Advances in Model-Oriented Design and Analysis*, Springer: 91-98.

Hotelling 1944 Some improvements in weighing and other experimental techniques. *Ann. Math. Statist.* 15(3): 297-306.

Huang Y, Gilmour SG, Mylona K and Goos P 2019 Optimal design of experiments for non-linear response surface models. *J. Royal Stat. Soc. Ser. C: Appl. Stat.* 68(3): 623-640.

Kaveh A and Talatahari S 2010 Imperialist competitive algorithm for engineering design problems. *Asian J. Civil Eng.* 11(6): 675-697.

Kiefer J 1959 Optimum experimental designs (with discussion). *J. R. Statist. Soc. B* 21: 272-319.

Mood A 1946 On Hotelling's weighing problem. *Ann. Math. Stat*. 17: 432-446.

Rodríguez G 2007 Lecture Notes on Generalized Linear Models. *URL:http://data.princeton.edu/wws509/notes/*

Russell KG, Woods DC, Lewis SM and Eccleston JA 2009 D-optimal designs for Poisson regression models. *Statistica Sinica* 19(2): 721-730.

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.

Wald A 1943 On the efficient design of statistical investigations. *Ann. Math. Stat.* 14: 134-140.

Wang Y, Myers RH and Ye K 2006 D-optimal designs for Poisson regression models. *J. Stat. Plan. Infer.* 136(8): 2831-2845.

Yang M, Zhang B and Huang S 2011 Optimal designs for generalized linear models with multiple design variables. *Statistica Sinica* 21(3): 1415-1430.