International Journal of Statistics and Probability; Vol. 9, No. 2; March 2020
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

D-optimal Design in Linear Model With Different Heteroscedasticity Structures

BODUNWA, O. K.[1], FASORANBAKU, O. A.[1]

[1]Department of Statistics, Federal University of Technology, Akure

Correspondence: BODUNWA, O. K, Department of Statistics, Federal University of Technology, Akure. E-mail: okbodunwa@futa.edu.ng

Received: January 6, 2020 Accepted: February 21, 2020 Online Published: February 26, 2020

doi:10.5539/ijsp.v9n2p7 URL: https://doi.org/10.5539/ijsp.v9n2p7

Abstract

In this paper, we developed D-optimal design in linear model with two explanatory variables in the presence of heteroscedasticity. A sequential method of getting D-optimal design was adopted. Two different structures were used based on the literatures; it was found that the optimal design takes the extreme values of the design region. The results of simulated data was justified with real life data from the kinematic viscosity of a lubricant, in stokes, as a function of temperature and pressure which was used as discussed in Linssen (1975). The relative efficiency of other designs with respect to D-optimal designs was determined. Three correction methods was adopted from weighted least square method for heteroscedasticity problem, it was found that the correction method tagged HCW1 performed better.

Keywords: D-optimal design, Heteroscedasticity, experimental design, sequential method, correction measure

1. Introduction

Experimentation is the process of planning a study to meet specified objectives which constitutes a foundation of the empirical sciences (Zhu, 2012). One major advantage of experiment is its ability to control the experimental conditions; as well as to determine the variables to include in a study (FackleFornius, 2008). Since the introduction of experimental design principle in the first half of the 1930, optimal experimental designs have been gaining attention and had become useful tools among researchers in various fields (Atkinson and Donev, 1992; Atkinson, 1996; Atkinson, Donev and Tobias, 2007; Berger and Wong, 2009). There are various design criteria, D-optimality has been the most frequently used; and often performs better than other criteria (Zocchi and Atkinson, 1999; Atkinson et al., 2007). Hence, the D-optimality has become one of the most popular criteria which involve designs that minimize the generalized variance of the parameter vector. The D-optimal designs seek to minimize $|(\mathrm{X}'\mathrm{X})^{-1}|$ (dispersion matrix) or equivalently maximise the determinant of the information matrix $(\mathrm{X}'\mathrm{X})$ of the design through some forms of statistical modeling such as regression model. One of the important assumptions of the standard regression model is that the variance of the error terms (disturbance term, u_i) must be equal across the observations which is refers to as homoscedastic with the model $y = x\beta + u_i$ where $[E(u_i^2) = \sigma^2 \quad i = 1,2,\cdots,n]$. However, in real life situations, this assumption is often violated and the variances of the error terms are not the same. The condition where error terms have different variances is termed heteroscedasticity $[E(u_i^2) = \sigma_i^2 \quad i = 1,2,\cdots,n]$ that is, unequal variance across the observations (Lambert, 2013; Knaub, 2017). Heteroscedasticity, which is often referred to as a "problem" that needs to be "solved" or "corrected" is the change in variance of predicted y, given different values of the independent variables (Knaub, 2011, 2017). The aim of this research work is to examine D- optimal Designs with different heteroscedastcity Structures and the objectives are to construct D-optimal design with different heteroscedasticity structures, to obtain the relative efficiencies of other designs with respect to D-optimal design, to determine the heteroscedasticity correction measure that will produce the most efficient D-optimal design in the different structures, determining the relative efficiencies of the parameters of the D-Optimal design model and to establish the best heteroscedasticity correction measure to achieve the most Efficient Parameter Estimation for D-Optimal Design.

Yan and Raymond (2001) presented D-optimal designs for two- variable logistic regression models where two-variable were fitted in the logistic regression models.Jafari (2013) found locally D-optimal design for a logit model in discrete choice experiment where there are many alternative set for people to make their choice using D-optimal design for the combination of the level of attributes to create alternatives. Jafari, *et.al*,(2014) worked on D-optimal design for logistic regression model with three independent variables; they obtained a locally D-optimal design for several specific states, presented certain designs with different points and calculated the subject optimality based on space of the parameters.

Jafari and Maram (2015) explored the notion of Bayesian D-optimal design for logistic regression model with exponential distribution for random intercept and obtained Bayesian D-optimal design; the method to maximize the Bayesian D-optimal criterion which is a function of the quasi- information matrix that depends on the unknown parameters of the model.

Jesús López-Fidalgo and Garcet-Rodríguez, (2004) considered the problem of constructing optimal designs for regression models when the design space is a product space and some of the variables are not under the control of the practitioner.Zhide and Douglas (2004) found locally D-optimal designs for multistage models and heteroscedastic polynomial regression model where they considered the construction of locally D-optimal designs for non-linear, multistage model in which one observes a binary response variable.Gaviriaa and López-Rósb (2014) worked on locally D-optimal designs with Heteroscedasticity: a comparison between two methodologies, it was found that the optimal design point takes the extreme values for both methods. These prior studies were more particular about the construction of the optimal designs with different models under some assumptions of the explanatory variables. In this study, construction of D-optimal designs in linear model with two explanatory variables in which there is a problem of heteroscedasticity in the model were examined. Different structures were used and the effects were also found on the optimal design.

2. Material and Method

2.1 Simulation Study

Starting with a linear regression model of the form (2.1)

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i \tag{2.1}$$

Where e_i is the error term which is a stochastic term assumed to be normally distributed with mean zero and variance σ_i^2 i.e. $e_i \sim N(0,\ \sigma_i^2)$.These $x_i s$ are fixed independently variables and y_i is the dependent variable and β_i are parameters that are known. The generations of the data used for independent variables are random variables that are normally distributed

$$x_1 = \left((1 - K\hat{}2)\hat{}0.5\right) * E_1 + K * E_2 \tag{2.2}$$

$$x_2 = \left((1 - K\hat{}2)\hat{}0.5\right) * E_2 + K * E_2 \tag{2.3}$$

Where K is the correlation between the explanatory variables, $E_1\ and\ E_2$ are the independent standard normal distribution with mean zero and the unit variance. The response variable was therefore obtained with equation

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \tag{2.4}$$

Where $e_i = Z_i\sqrt{Var(e_i)}$, $Z_i \sim N(0,1)$ $\quad i = 1,\ 2.$

$e_1 \sim N(0, X_{2i}^2)$ (Park, 1966, White, 1980, Guajarati *et. al* 2012)

$e_i \sim N\left(0, Exp(x_2^2)\right)$ (Box and Hill, 1974, Harvey, 1976)

The $Var(e_i)$ took any of the structures in equations 2 and 3. The simulations were carried out in one thousand times (1000) at eight sample sizes of 10, 20, 30, 40, 50, 100, 250 and 500.

In order to correct the heteroscedasticity problem with the selected structures, the weighted least square methods was adopted and the $log\hat{e}_i^2\ was$ regress $on\ (x_1, x_2)$ to have Heteroscedasticity Correction Weighted 1 (HCW1), $log\hat{e}_i^2\ on\ (x_1,\ x_2, x_1^2, x_2^2)$ to have Heteroscedasticity Correction Weighted 2 (HCW2) and $log\hat{e}_i^2\ on\ (x_1,\ \ x_2, x_1^2,\ \ x_2^2,\ x_1 x_2)$ to have Heteroscedasticity Correction Weighted 3 (HCW3).

2.2 Construction of D-optimal Design

There are several methods at hand on the practices of determining the optimal design. These include algorithms, sequential, analytical, numerical and graphical methods, used separately or in combinations. There is no method that is generally favorable; it depends on the problem at hand. The method selected in this research work is sequential method of getting D-optimal design; we find the D-optimal design for model with different variance structure of the error term was essentially obtained. For the model (2.1) used in this study, the number of p is 3. Therefore the partial derivative for the model is

$$f'(x_i) = (1,\ x_1,\ x_2) \tag{2.5}$$

The information matrix is now

$$M(\xi) = \sum w_i f(x_i)\, f'(x_i) \tag{2.6}$$

Beginning with p-point design, we get 3×3 design matrix of the form

$$X_3 = \begin{bmatrix} 1.000000 & 1.000000 & 0.322035 \\ 1.000000 & -0.494439 & 0.413935 \\ 1.000000 & -0.235592 & -0.026634 \end{bmatrix} \quad (2.7)$$

It should be noted that the procedure requires a sufficient number of observations because we have to ensure that the inverse $|X_N' X_N|^{-1}$ exist. A simple condition that will guarantee the inverse exists is to have the number of different design points greater than or equal to the number of parameters, that is $N \geq p$

The design points are selected within the range of $-1 \leq x \leq 1$ for the variables. The largest $s(x_a, \xi)$ is found for $x_1 = 1.00000\ and x_2 = -1.000000$, so these design points were added to design matrix X_3 and the design matrix is now

$$X_4 = \begin{bmatrix} 1.000000 & 1.000000 & 0.322035 \\ 1.000000 & -0.494439 & 0.413935 \\ 1.000000 & -0.235592 & -0.026634 \\ 1.000000 & 1.000000 & -1.000000 \end{bmatrix} \quad (2.8)$$

The iteration continued until the condition for getting optimal design was reached. The maximum $s(x_i, \xi)$ value decreases as N increases, according to the general equivalence theorem (Kiefer and Wolfowitz, 1960), a D-optimal design satisfies the condition that $s(x_a, \xi) \leq p$.

2.3 Relative Efficiencies of D-optimal to Other Designs

The Efficiency of D-optimal design ξ_D with respect to the other design is

$$D_{eff} = \left(\frac{|M(\xi)|}{|M(\xi_D|)}\right)^{1/p} \quad (2.9)$$

Where p is the number of parameters of the model and $M(\xi)$ denotes the information matrix of the design ξ which is another design different from D-optimal design. Relative efficiencies of the parameters of the D-optimal design and non optimal designs models were also done to establish the result of D-optimal designs point. The design points for all the structures were obtained with respect to the probability, number of iteration, the standardized variance.

2.4 Most Efficient Correction Method

The best correction method among the one named HCW1, HCW2 and HCW3 was determined. This was done by calculating the variances for the probabilities of the D-optimal designs taking the design points as x and the probabilities as $f(x)$. The minimum variances were selected for the structures for all the sample sizes and the method that has highest values was chosen to be the most efficient.

2.5 Real Life Application

Construction of D-optimal design in the presence of heteroscedasticity for the model (1) was applied to a real life data, a secondary data from the kinematic viscosity of a lubricant, in stokes, as a function of temperature(o_C), and pressure in atmospheres (atm), was used as discussed in Linssen (1975) where y is predicted In (viscosity), x_1 is temperature, and x_2 is pressure to justify the simulated data.

3. Result and Discussion

In this work, D-optimal designs with two different heteroscedasticity structures were constructed when there is no heteroscedasticity (No H) and when there is (HR). It was generally found that the D-optimal designs take the extreme values of the response variables which follow uniform distribution of the experimental units

Table 3.1. D-Optimal Designs for the Structures

Structures		(-1, -1)	(-1, 1)	(1, -1)	(1, 1)
$\sigma^2 X_{2i}^2$	No H	44(0.25143)	44(0.25143)	44(0.25143)	43(0.24571)
	HR	28(0.24138)	30(0.25862)	295(0.25000)	29(0.25000)
$\sigma^2 Exp(x_2^2)$	No H	44(0.25143)	44(0.25143)	43(0.24571)	44(0.25143)
	HR	22(0.23656)	24(0.25806)	24(0.25806)	23(0.24731)

Table 3.1 presents the construction of the D-optimal when there is no heteroscedasticity and when there is heteroscedasticity for the error structures. It can be seen that the D-optimal designs when there is no heteroscedasticity for the two structures were same reason being that the error term have equal variance. The optimal designs even though the model has three parameters the design consists four points which are the extreme points of the regression range.From the table, it can been seen that

$$\xi^* = \begin{Bmatrix} (-1,-1) & (-1,1) & (1,-1) & (1,1) \\ 0.24138 & 0.258620.25000 & 0.250000 \end{Bmatrix} \tag{3.1}$$

if there are 116 experimental units, 28 should be allocated to when $x_1 = -1$ and$x_2 = -1$, 30 should be for when $x_1 = -1$ and$x_2 = 1$. In the same vein, 29 should be allocated to when $x_1 = 1$ and $x_2 = -1$ and when$x_1 = 1$ and$x_2 = 1$.

Considering D-optimal design for the second structure,

$$\xi^* = \begin{Bmatrix} (-1,-1) & (-1,1) & (1,-1) & (1,1) \\ 0.23656 & 0.258060.25806 & 0.24731 \end{Bmatrix} \tag{3.2}$$

Equation shows that if there are 93 experimental units, 22 should be allocated to when $x_1 = -1$ and $x_2 = -1$, 24 should be for when $x_1 = -1$ and $x_2 = 1$and when$x_1 = 1$ and $x_2 = -1$, 23 for when $x_1 = 1$ and $x_2 = 1$.

Table 3.2. D-optimal Designs for the real life data

	(-1,-1)	**(-1,1)**	**(1,-1)**	**(1,1)**
HR	16(0.30000)	11(0.20000)	16(0.30000)	11(0.20000)

The results still revealed that the D-optimal design for the real life data presented above affirmed the result from simulated data in the sense that the design point takes the extreme values of the design region.

The relative efficiencies of D-optimal design with respect to other designs that are not optimal using the same method of construction of D-optimal design from the starting design matrix of point 4 is given below for the structures.

Table 3.3. Relative Efficiency Table

$\sigma^2 X_{2i}^2$		$\sigma^2 Exp(x_2^2)$	
No of Iteration	D-efficiency	No of Iteration	D-efficiency
4	0.0019	4	0.0043
5	0.0225	5	0.0329
6	0.0331	6	0.0453
⋮	⋮	⋮	⋮
114	0.9829	91	0.9788
115	0.9914	92	0.9894

Table 3.3 shows that the D-optimal design has close efficiency to other design especially the one closed to the design point meaning that the closer the D-efficient to one, the better. The no of iteration for D-optimal design for the first structure is 116 and for the second structure 93. Next table present the D-efficiency of the real life data.

Table 3.4. Relative Efficiencies of other Designs for real life data

I	$\boldsymbol{D_{eff}}$
4	0.002128
5	0.003511
6	0.004728
⋮	⋮
901	0.9869
902	0.9931

To determine the best correction method, the variances of the probability in the design point of the D-optimal design were calculated using different sample sizes. The best method was chosen on the basis of the one with minimum variance. Table 3.5 presented the variances of design points.

Table 3.5. Determination of the best Correction Method

Forms	Correction Methods	Sample size 10	20	30	40	50	100	250	500
$\sigma^2 X_{2i}^2$	HCW1	1.24996	1.26050	1.24407	**1.23217**	1.24290	**1.24434**	**1.24407**	1.25584
	HCW2	**1.24978**	**1.25912**	1.25000	1.24386	**1.24113**	1.25762	1.25584	**1.24172**
	HCW3	1.24995	1.25969	**1.24362**	1.24386	1.25718	1.25000	1.25598	1.24223
$\sigma^2 Exp(x_2^2)$	HCW1	**1.25774**	1.26146	1.24362	**1.23030**	1.24362	**1.25534**	**1.24481**	**1.25628**
	HCW2	1.25786	**1.25868**	1.23777	1.24401	**1.24144**	1.25786	1.25546	1.25739
	HCW3	1.27398	1.25899	**1.21102**	1.2386	1.25899	1.25718	1.25523	1.25762

From the table, number of appearance of minimum variance values in HCW1 is more than the other two. Therefore HCW1 is assumed to be performing better.

4. Conclusion

In the study, constructions of D-optimal designs in the presence of Heteroscedasticity for two different structures were considered with when there is no Heteroscedasticity in the data.

It was generally found that the D-optimal designs take the extreme values of the response variables which follow uniform distribution of the experimental units which can be interpreted as taking the least and the highest values of the explanatory variables in order to get best output through the response variable. To verify the above findings, a set of real life data (secondary data) was used and the design points for D-optimal designs were same with simulated data.

The relative efficiencies of other designs under different Heteroscedasticity structures were found to prove the strength of the design. Determination of the best correction method was also found. This was achieved by comparing the variances of the selected correction methods with respect to sample sizes for all the structures used in the study. It was found that the correction method with minimum variance that showed the efficiency of the method represented by (HCW1) which was done by regressing $log\hat{e}_i^2$ on the linear combinations of $x_1 \, and \, x_2$performed better than the remaining two.

References

Atkinson, A. C., Donev, A. N., & Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*. Oxford University Press.

Atkinson, A. C. (1996). The usefulness of optimum experimental designs.*Journal of the Royal Statistical Society. Series B, 58*, 59-76.https://doi.org/10.1111/j.2517-6161.1996.tb02067.x

Atkinson, A. C., & Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford University Press, Oxford.

Berger, P. F., & Wong, K. W. (2009). *An Introduction to optimal designs for social and Biomedical research.*A John Wiley & Sons, Ltd. Publication.https://doi.org/10.1002/9780470746912

Box, G. E. P., & Hill, W. J. (1974). Correction Inhomogenity of Variance with Power Transformation Weighting. *Technometrics, 16*(3), 385-389. https://doi.org/10.1080/00401706.1974.10489207

Fackle, F. E. (2008). *Optimal Design of Experiments for the Quadratic Logistic Model.* A Thesis submitted to the Department of Statistics, Stockholm University, Stockholm, in partial fulfillment of Doctor of Philosophy in Statistics.

Kiefer, J., & Wolfowitz, J. (1960). The Equivalence of Two Extremum Problems. *Canad. J. Math., 12*, 363-366. https://doi.org/10.4153/CJM-1960-030-4

Knaub, J. R. J. (2011). Ken Brewer and the Coefficient of Heteroscedasticity as Used in Sample Survey Inference.*Pakistan Journal of Statistics*, *27*(4), 397-406.

Knaub, J. J. R. (2017). Essential Heteroscedasticity. Retrieved from https://www.researchgate.net/publication/320853387_Essential_Heteroscedasticity

Gaviriaa, J. A., &L ópez-R ósb, V. I. (2014). Locally D-Optimal Designs with Heteroscedasticity: A Comparison between Two Methodologies. *RevistaColombiana de Estad ística, 37*(1), 95-110. https://doi.org/10.15446/rce.v37n1.44360

Gujarati, N. D., Porter, C. D., & Gunasekar, S. (2012). "Basic Ecocnometric" (Fifth Edition) New Delhi: Tata

McGraw-Hill.

Jafari, H., & Maram. (2015). Bayesian D-optimal design for Logistic Regression model with Exponential distribution for random intercept. *Journal of Statistical Computation and Simulation.*

Jafari, H., Khazai, S., & Khaki, Y. (2014). D-optimal design for logistic regression model withthree independent variables.*Journals of Asian Scientific Research, 4*(3), 120-124.

Lambert, B. (2013). Heteroscedasticity Summary, June 3, 2013, YouTube. Retrieved fromhttps://youtu.be/zRklTsY9w9c

Linssen, H. N. (1975). Nonlinearity measures: a case study, Statist. *Neerland, 29*, 93-99. https://doi.org/10.1111/j.1467-9574.1975.tb00253.x

López-Fidalgo, J., & Garcet-Rodríguez, S. A. (2004). Optimal experimental designs when some independent variables are not subject to control. *Journal of the American Statistical Association*, *99*(468), 1190-1199. https://doi.org/10.1198/016214504000001736

Park, R. E. (1966). Estimation with Heteroscedastic Error terms. *Econometrica, 34*, 888-892. https://doi.org/10.2307/1910108

White, H. (1980). A Heteroscedastic-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica, 48*, 817-838. https://doi.org/10.2307/1912934

Zhu, C. (2012). *Construction of Optimal Designs in Polynomial Regression Models*. A Thesis submitted to the Faculty of Graduate Studies of The University of Manitoba in Partial Fulfillment of the Requirements for the Degree of Master of Science Department of Statistics University of Manitoba Winnipeg, Manitoba, Canada.

Zocchi, S. S., & Atkinson, A. C. (1999). Optimum experimental designs for multinomial logistic models. *Biometrics, 55*, 437-444. https://doi.org/10.1111/j.0006-341X.1999.00437.x