HETEROSCEDASTIC D-OPTIMAL DESIGN IN REGRESSION MODELS WITH APPLICATION IN KINEMATIC VISCOSITY DATA

Bodunwa O.K*, Fasoranbaku O.Aand Ajiboye. A. S.

Department of Statistics,
Federal University of Technology,
Corresponding Author's e-mail;okbodunwa@futa.edu.ng

Abstract
In real life situations, the assumption of homogeneity is often violated and the variances of the error terms are not the same, heteroscedastic problem. In this work, D-optimality criterion which is an experimental design was used when there is heterocedsasticity in the data set of kinematic viscosity and when the data had been corrected using different methods for correction thereby making the variance of the error structure to be equal. Comparing the resultfor the two regression models used, when there is heteroscedasticity and when it has been corrected, the variances and the determinants of dispersion matrices shows that D-optimal design when the data set has been corrected is more efficient than when there is heteroscedastic .

Key words: D-optimal, Heteroscedasticity, Experimental Design, Regression model

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 theexperimental conditions as well as to determine the variables to include in a study (Fackle Fornius, 2008). Since the introduction of experimental design principle in the first half of the 1930, optimal experimental designs have been gaining attentionand 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 involves designs that minimize the generalized variance of the parameter vector.The D-optimal designs seek to minimize$|(X'X)^{-1}|$(dispersion matrix) or equivalently maximize the determinant of the information matrix $(X'X)$of the design through some forms of statistical modeling such as regression model. The information matrix (also called Fisher information matrix)measures the amount of information that random variable Xaffects an unknown parameter θ of a distribution.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 and this is refers to as homoscedastic $[E(u_i^2) = \sigma^2 i = 1,2,\cdots,n]$. 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 $[E(u_i^2) = \sigma_i^2 i = 1,2,\cdots,n]$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).This study therefore, adoptedD-optimal designs otherwise known as D-optimality criterion in regression model with heteroscedasticity error terms to provide the most accurate estimates of model parameters using Kinematic Viscosity Data.

LITERATURE REVIEW

Atkinson (1996) gave a compelling account of the usefulness of optimal design and their potential application to other fields. Berger and Wong (2009) showed interest in application of optimal design in different disciplines.Mannarswamy,*et.al*, (2009) derived D-optimal experimental designs for three common adsorption isotherm models: the 2-parameter Freundlich, the 2-parameter Langmuir, and the 3-parameter Langmuir. For each of these liquid-solid adsorption models, the D-optimal design equations were derived from the information matrix and then solved numerically to determine the specific design values as a function of the model parameters. It was determined that the D-optimal designs for all three isotherms were independent of the proportionality parameter in the model and the design optimality was verified using the general equivalence theorem. Fang (2002) considered D-optimal designs for polynomial regression models with low-degree terms being missed by applying the theory of canonical moments. It turns out that the optimal design places equal weight on each of the zeros of some Jacobi polynomial when the number of unknown parameters in the model is even.Wang, et.al, (2006) worked on D-optimal designs for Poisson regression models. For the one-variable first-order Poisson regression model, it has been found that the D-optimal design in terms of effective dose levels is independent of the model parameters. Habib (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.Habib, 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. Gaviriaa and López-Ríosb (2014) worked on Locally D-Optimal Designs with heteroscedasticity. In their work, two methodologies were considered in a non linear model with one explanatory variable. They used variance modeling methodology and Box-Cox transformation to get D-optimal design.

MATERIALS AND METHOD

The data usedin this study is a secondary data on the kinematic viscosity of a lubricant (response variable) in stokes as a function of temperature (o_C) and the pressure in atmosphere (atm), were obtained by Linssen (1975). The data setwhich was tested and found to be violating the assumption of homoscedasticity was codedbetween$-1 \leq x \leq 1$.

The Linear model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i$$

where y_i is the viscosity, x_1 is the temperature and $x_2$is the pressure, the parameters are β_i and $e_i$is the error term.

The polynomial model was also used and we found D-optimal design for each modelusing sequential method when there is heteroscedasticity and when it has been corrected.The sequential method for getting D-optimal design was used to achieve the results in this work

The sequential method of getting D-optimal design 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 exist 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. Wherever the maximum value of the Standardized variance $s(x_a, \xi)$ is found will be added to the next design matrix. The process continued until the condition for getting optimal design was reached. The maximum $s(x_a, \xi)$ value decreases as N increases,according to the general equivalence theorem (Kiefer and Wolfowitz, 1960), a D-optimal design satisfies the condition that $s(x_a, \xi) \leq p$.

The coded values for x_1 and x_2 uses $d_j = \frac{x_i - D}{x_{max} - D}$ where $D = x_{min} - \frac{x_{max} - x_{min}}{2}$

Heteroscedasticity was corrected using the following steps

i. Run the OLS of the model and obtain the estimated residuals
ii. Obtain the log of the squared residuals
iii. Correction methods are

Method 1 regress $log\hat{e}_i^2 \ on \ (x_1, x_2)$

Method 2 regress $log\hat{e}_i^2 \ on \ (x_1,\ x_2, x_1^2, x_2^2)$

Method 3 regress$log\hat{e}_i^2 \ on \ (x_1,\ x_2, x_1^2,\ x_2^2,\ x_1x_2)$

RESULTS AND DISCUSSION

The linear (first) model is

$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i (1)$

The partial derivative for the model is

$$f'(x_i) = (1,\ x_1,\ x_2) \qquad (2)$$

The information matrix is now

$$M(\xi) = \sum w_i \, f(x_i) \; f'(x_i) \qquad (3)$$

The corresponding 3×3 design matrix for the model is

$$X_3 = \begin{bmatrix} 1.000000 & 1.000000 & 0.322035 \\ 1.000000 & -0.494439 & 0.413935 \\ 1.000000 & -0.235592 & -0.026634 \end{bmatrix} \qquad (4)$$

The maximum$s(x_a, \xi)$ 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

$$X_4 = \begin{bmatrix} 1.000000 & 1.000000 & 0.322035 \\ 1.000000 & -0.494439 & 0.413935 \\ 1.000000 & -0.235592 & -0.026634 \\ 1.000000 & 1.000000 & -1.000000 \end{bmatrix} \qquad (5)$$

Table 1 shows the D-optimal design when there is heteroscedasticity. It means that if there are 100 experimental units, 20 should be allocated to when $x_1 = -1$ and$x_2 = 1$, also when $x_1 = 1$ and $x_2 = 1$. In the same vein, 30 should be allocated to when $x_1 = -1$ and$x_2 = -1$, also when $x_1 = 1$ and $x_2 = -1$

Table1: Sequential construction of a D-optimal design for the linear (first) model

N	x_{1N+1}	x_{2N+1}	$s(x_a, \xi)$	D_{eff}
3	1.0000	−1.0000	49.5896	0.000819
4	−1.0000	−1.0000	13.9742	0.002128
5	−1.0000	1.0000	7.2029	0.003511
6	1.0000	1.0000	6.0575	0.004728
7	−1.0000	−1.0000	4.3792	0.005966
8	1.0000	−1.0000	4.3516	0.007015
9	−1.0000	1.0000	4.2274	0.008169
10	1.0000	1.0000	3.7240	0.00931
903	−1.0000	−1.0000	3.009	1

The D-optimal design for the model is

$$\xi^* = \begin{Bmatrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) \\ \frac{2}{10} & \frac{2}{10} & \frac{3}{10} & \frac{3}{10} \end{Bmatrix} \qquad (6)$$

Comparison for D-optimal designs when there is heteroscedaticicty and when it has been corrected was presented in Table 2 for linear model. The Mean and Variance for these were shown. The correction methods used here were three and this was done for polynomial model

Table 2: D-optimal designs for model 1 {$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i$(Linear Model)}

Model	D-optimal Designs	Mean	Variance
Uncorrected	$\left\{\begin{matrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) \\ \frac{2}{10} & \frac{2}{10} & \frac{3}{10} & \frac{3}{10} \end{matrix}\right\}$	**2.7**	**8.5**
Corrected 1	$\left\{\begin{matrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) \\ \frac{5}{20} & \frac{5}{20} & \frac{5}{20} & \frac{5}{20} \end{matrix}\right\}$	**2.5**	**7.5**
2	$\left\{\begin{matrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) \\ \frac{5}{20} & \frac{6}{20} & \frac{4}{20} & \frac{5}{20} \end{matrix}\right\}$	**2.45**	**7.25**
3	$\left\{\begin{matrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) \\ \frac{5}{20} & \frac{5}{20} & \frac{5}{20} & \frac{5}{20} \end{matrix}\right\}$	**2.5**	**2.5**

From Table 2, the mean and variance for the D-Optimal design when there is heteroscedasticity and it has been corrected were shown. The results show that the variance of the D-Optimal when there is heteroscedasticity is higher than when it has been corrected, which show the efficiency of the corrected one. Similarly, Table 3 also show same results for the second model used.

The polynomial model considered in this work is $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + e_i$

Table 3: D-optimal designs for model 2 (Polynomial Model)

Model	D-optimal Designs	Mean	Variance
Uncorrected	$\begin{Bmatrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) & (-0.2356,1) & (-0.2356,-1) \\ \frac{3}{20} & \frac{3}{20} & \frac{3}{20} & \frac{4}{20} & \frac{3}{20} & \frac{4}{20} \end{Bmatrix}$	**3.65**	**16.25**
Corrected 1	$\begin{Bmatrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) & (-0.2356,1) & (-0.2356,-1) \\ \frac{4}{20} & \frac{4}{20} & \frac{2}{20} & \frac{3}{20} & \frac{4}{20} & \frac{3}{20} \end{Bmatrix}$	**3.4**	**14.3**
2	$\begin{Bmatrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) & (-0.2356,1) & (-0.2356,-1) \\ \frac{4}{20} & \frac{4}{20} & \frac{2}{20} & \frac{3}{20} & \frac{4}{20} & \frac{3}{20} \end{Bmatrix}$	**3.4**	**14.7**
3	$\begin{Bmatrix} (-1,1) & (1,1) & (-1,-1) & (1,-1) & (-0.2356,1) & (-0.2356,-1) \\ \frac{4}{20} & \frac{4}{20} & \frac{4}{20} & \frac{4}{20} & \frac{2}{20} & \frac{2}{20} \end{Bmatrix}$	**3.1**	**12.1**

Table 3 below shows the D-optimal design and the comparison is based on the mean and the variance of the design knowing fully well that for a good estimator, efficiency must be ascertained which is relevant to minimum variance. Therefore, the D-optimal design when there is no heteroscedasticity is the best in experimental design. Table 4 further established the effect of heteroscedasticity on D-optimal design for itsminimizes the determinant of dispersion matrix or maximize the information matrix, meaning that the smaller the determinant, the better the design. Model 1 can be considered to have performed best relative to this fact

Table 4: Comparison of D-optimal designs in relative to the determinants of dispersion matrices

Model	heteroscedasticity	No Heteroscedasticity		
		Correction Method 1	2	3
Model 1	$3.331295e^{-11}$	$3.331158e^{-11}$	$3.329784e^{-11}$	$3.323053e^{-11}$
Model 2	$2.498338e^{-11}$	$2.498807e^{-11}$	$2.499842e^{-11}$	$2.4987226e^{-11}$

CONCLUSION

Generally, in regression model presence of heteroscedasticity makes the estimator unbiased but not efficient when ordinary least squares (OLS) approach is used. In view of this, finding D-optimal design when Heteroscedasticity has been corrected is better than when left uncorrected. In optimality criterion too, presence of this phenomenon also affect the D-optimal design that an experimenter wishes to achieve.It is recommended that when finding D-optimal design, heteroscedasticity should be corrected especially when dispersion matrix is involved in the method to get the optimal design.

REFERENCES

Atkinson, A. C., Donev, A. N. (1992). Optimum Experimental Designs. Oxford Statistical Science Series-8, Oxford University Press.

Atkinson, A. C., Donev, A. N. and 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.

Atkinson, A.C. and Cook, R.D. (1995). D-optimum designs for heteroscedastic linear models. Journal of the American Statistical Association, 90, 204-212.

Atkinson, A.C. and Donev, A. N. (1996). Experimental designs optimally balanced for trend. Technometric, 38,333-341.

Berger, M. P. F. and Wong, W. K. (2009). An introduction to optimal designs for social and biomedical research. John Wiley & Sons.

Dette, H. and Muller, W. G. (2012). Optimal designs for regression models with a constant coefficient of variation AMS Subject Classification: 62K05in non-linear regression. Statistica Sinica, 21 1557-1570.

Dette, H. Melas, V. B. and Shpilev, P. (2008). Optimal designs for estimating the derivative

Fackle Fornius, 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.

Fisher, R.A. (1930). The Design of Experiments, Oliver and Boyd, Edinburgh.

Gaviria J.A. and López-Ríosb V.I. (2014). Locally D-Optimal Designs with Heteroscedasticity: A Comparison between Two Methodologies. Revista Colombiana de Estadística. 37 (1): 95-110.

Habib, J. (2013). On Locally Optimal Criterion for a Logit Model with Random Parameters. *Applied Mathematics and Information Sciences* 7 (2) 732-728

Jafari and Maram (2015). Bayesian D-optimal design for logistic regression model with exponential distribution for random intercept. Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2015.1087525

Jie Yang, Liping Tong and Abhyuday Mandal (2015).D-optimal Designs with Ordered CategoricalData. Statistica Sinica · DOI: 10.5705/ss.202016.0210 ·

Kiefer, J. and Wolfowits, J. (1959). Optimum designs in regression problems. The annals of Mathematical Statistics, 30,271-94.

Klein A.G., Gerhard C., Büchner R.D., Diestel S. and Schermelleh-Engel K (2016). “The detection of heteroscedasticity in regression models for psychological data” Psychological Test and Assessment Modeling, 58 (4), 567-592.

Klein A.G., Gerhard C., Büchner R.D., Diestel S. and Schermelleh-Engel K (2016) “The detection of heteroscedasticity in regression models for psychological data” Psychological Test and Assessment Modeling, 58 (4), 567-592

Knaub J.R., Jr. (2017). “Essential Heteroscedasticity,” unpublished research, https://www.researchgate.net/publication/320853387_Essential_Heteroscedasticiy

KnaubJ.R., Jr. (2011). “Ken Brewer and the Coefficient of Heteroscedasticity as Used in Sample Survey Inference,” Pakistan Journal of Statistics, 27 (4): 397-406.

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

Lopez-Fidalgo, J. Rivas-Lopez, M.J. and Del Campo, R. (2009). Optimal designs for Cox regression. Statistica neerlandica, 63 (2): 135-148 doi.org/10.111/j.1467-9574.2009.00415

Mannarswamy, A., Munson-McGee, S. H., Steiner, R., & Andersen, P. K. (2009). D-optimal experimental designs for Freundlich and Langmuir adsorption isotherms. *Chemometrics and Intelligent Laboratory Systems*, *97*(2), 146–151.

Martijin P.F. Berger, Weng Kee Wong (2009). An introduction to optimal designs for social and biomedical research.

Melas, V. (2006). Functional Approach to Optimal Experimental Design. Lecture Notes in Statistics 184, Springer-Verlag, New York.

Montgomery, D. C. (2000). Design and Analysis of Experiments, 5th Edition. Wiley, New York.

Montgomery, D.C; Peck, E. A. and Vining, G. G. (2001). Introduction to Linear Regression Analysis. Wiley, New York.

Timothy E.O'Brien and Gerald M. Funk (2003). "a gentle introduction to optimal design for regression models" the American statistician. 57 (4): 265-267.

Ullah M. A and Imdadullah, M.(2011). Impact Analysis for Regression Models with Heteroscedastic Errors Pakistan Journal of Social Sciences (PJSS)31 (2). 379-394

Wang, Y., Myers, R. H., Smith, E. P., & Ye, K. (2006). D-optimal designs for Poisson regression models. *Journal of Statistical Planning and Inference*, *136*(8), 2831–2845.

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. and Atkinson A.C. (1999). Optimum experimental designs for multinomial logistic models, Biometrics, 55, 437-444.