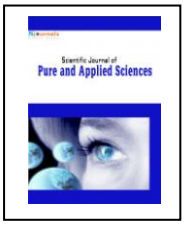
Sjournals
Scientific Journals
Scientific Journal of Pure and Applied Sciences (2015) 4(1) 28-33
ISSN 2322-2956
doi: 10.14196/sjpas.v4i1.1777

Contents lists available at Sjournals
Scientific Journal of
Pure and Applied Sciences
Journal homepage: www.Sjournals.com



Original article

Optimal design for count data

M. Abdulkabir[a,*], U. Anietie Edem[b], B. Latifat Kemi[b]
[a]*Postgraduate Student University of Ilorin, Ilorin, NIGERIA.*
[b]*Mathematics and Statistics Department Federal Polytechnic, Offa, Kwara State, NIGERIA.*

*Corresponding author; Postgraduate Student University of Ilorin, Ilorin, NIGERIA.

ARTICLE INFO

Article history,
Received 11 December 2014
Accepted 22 December 2014
Available online 29 December 2014

Keywords,
Generalized linear model (GLM)
Optimality criterions

ABSTRACT

Optimal designs for generalized linear models (GLM) have received increasing attention in recent years. Most of this research focuses on binary data model. This research extends to count data models. The aim and objectives of this research work to determine the appropriate generalized linear model (GLM) that is suitable for count data and identify a design that is best according to statistical optimality criteria, the data use for this research work are simulated data from R statistical package using uniform distribution with sample size 300. The simplest distribution use for modeling count data is Poisson distribution, quasi Poisson were carried out to test for over dispersion in the Poisson regression model and the formal way of dealing with over dispersion is negative binomial regression model, thus AIC was use to compare the two models, the Poisson regression model shows the best with minimum AIC. Furthermore optimal design were carried out using the optimality criterion that is the A and D optimality criterion, using design efficiency to compare the two (2) designs the optimality criterion with the highest efficiency is the best, thus D optimality criterion shows the best design.

1. Introduction

1.1. Optimal design

Optimal designs are class of experimental design that are optimal with according to some statistical optimality criterion. In the design of experiment for estimating statistical models, this design allows parameter to be estimate without bias and with minimum variance. An optimal design is considered one of the most important topics in the context of the experimental design; optimal design is the design that achieves some targets of our interest. The optimal designs are experimental designs that are generated based on a particular optimality criterion and are generally optimal only for a specific statistical model. Smith (1918) guest how optimality design experiments originated, he was one of the first to state a criterion and obtain optimal designs for regression problems. Many years later, Kiefer (1959) developed useful computational procedures for finding optimum designs in regression problems of statistical inference. An optimality criterion shows how good a design is, there are also designs that optimize the design space based on the decision framework. Some examples of such designs are A-, D-, E-optimal designs, and others. The idea of an optimal design is that statistical inference about the quantities of interest can be improved by appropriately selecting the values of the control variables. These values should be chosen such that variability of the estimator of the parameters is minimized, or certain risk is minimized. Because of the costs and/or other limitations on resources or time, efficient use of available resource for experimental design is critical. Numerous criteria have been developed to measure the performance of an experimental design based on the optimal design. Optimal design are also called optimum design

Optimal experimental design theory is a flexible approach used to design experiments. It is flexible because it can be used for any statistical model with any number of explanatory variables; both qualitative and quantitative, over any experimental design region and with any number of observations. When designing experiments using optimal design theory, the form of the model, whether it is linear or not, has important implications.

When the underlying model is linear, the optimal design process is not hindered by model parameters being unknown because they do not enter the optimality criteria. The covariance matrix of Y is not a function of the model parameters. As such, most of the work in experimental design has been based on models with a continuous response variable where the error term is assumed to be normally distributed with a constant variance. When the model is linear, there are many classical design techniques to choose from, such as factorial designs or response or response surface methodology, and the optimal design can be determined explicitly, Khuri et al. (2006). For this reason, in many situations, experiments are designed based on techniques for linear model even when this assumption is not valid. When the conditions for a linear model are not satisfied, such as in experiments with a binary (for example, defective/non defective) or count (for example, number of defectives) response, GLMs are appropriate. However, in this case, the optimal design process becomes complicated because it depends on the values of the model parameters. That is, the parameters values must be known to design an optimal experiment to estimate the parameters. The covariance matrix of Y depends on the value of the model parameters. This conundrum has hindered the development of theory associated with experimental design for GLMs.

The objectives of this study are to determine the appropriate generalized linear model (GLM) that is suitable for count data and to identify a design that is best according to statistical optimality criterion. The data used for the study were simulated data from uniform distribution that is X1~U (300, 0, 1.5) and X2 ~U (300, 0, 2.5) using R statistical software

2. Materials and methods

2.1. Count data

Count data are non negative integers, they represent the number of occurrence of an event within a fixed period, e.g. number of trade in a time interval, number of given disaster, number of crime on campus per semester e.t.c.

2.2. Poisson regression model

The simplest distribution used for modeling count data is the Poisson distribution, thus Poisson regression model is a special case of the generalized linear model (GLM) framework. The variance in the Poisson model is

identical to mean, thus the dispersion is fixed at theta given to be 1 and the variance function is V(μ)=μ. According to (McCullagh and Nelder 1989)
Coefficients:

Table 2.1
Poisson Regression Model.

	Estimate	Std. Error	z value	Pr(>|t|)
(Intercept)	0.96905	0.07513	12.898	< 2e-16 ***
x1	0.21014	0.05797	3.625	0.000289 ***
x2	0.41413	0.03533	11.723	< 2e-16 ***

(Dispersion parameter for Poisson family taken to be 1).
AIC: 1373.6.

All regressors are highly significant and the standard errors are appropriate. This will also be confirmed by the models that deal with over-dispersion (excess zeros) that is the quasi Poisson regression model.

2.3. Quasi-poisson regression model

The quasi Poisson model is estimated when there is presence of over dispersion or excess zeros in Poisson model thus the regressor for both quasi and Poisson model and the same AIC in model which are highly significant
Coefficients:

Table 2.2
Quasi-Poisson model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.96905	0.07765	12.480	< 2e-16 ***
x1	0.21014	0.05991	3.507	0.000522 ***
x2	0.41413	0.03651	11.343	< 2e-16 ***

(Dispersion parameter for quasi Poisson family taken to be 1.068064).
AIC: NA.

From the quasi Poisson model the estimated dispersion parameter were give as 1.068064 which is greater than 1 indicating that over-dispersion is present in the data. The result from quasi Poisson regression tests of the coefficients are the same as to the results obtained from the Poisson regression with standard errors, leading to the same conclusions as before.

2.3. Negative binomial regression model

Another way of modeling over dispersion count data is to assume a Negative binomial distribution for yi/xi which arises as a gamma mixture of Poisson distribution. (Nelder and Wedderburn 1972; McCullagh and Nelder 1989)
Coefficients:

Table 2.3
Negative Binomial Regression Model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.96962	0.07628	12.711	< 2e-16 ***
x1	0.20938	0.05909	3.544	0.000395 ***
x2	0.41413	0.03594	11.524	< 2e-16 ***

(Dispersion parameter for Negative Binomial (158.6768) family taken to be 1.
AIC: 1375.4.

2.4. Akaike information criterion (AIC)

AIC is a statistical measure of the likelihood of a model parameter for the complexity of the model. It is useful when comparing two or more models for data, which implies that all the data, must have the same independent

variables. The smaller the AIC the better fitted models of the parameter estimate. Comparing the models of Poisson regression model and negative binomial regression model the AIC for Poisson model is 1373.6 and the AIC for negative binomial is 1375.4 which implies the best model is Poisson regression model having the smallest AIC.

3. Results

A-optimality Criterion: Introduced by Chernoff (1953), showed the employed criterion of optimality which is the one that involves the use of Fisher's information matrix. An algebraic approach for constructing A–optimal design under generalized linear models was presented by Yang (2008). A–optimality is defined as: $\min_{x_i i=1,\dots n} trace\left(X^{1}X\right)^{-1}$ equivalently, minimizing the average variance of the parameter estimates. A efficiency of a design ξ is defined as:

$$A(\xi)=\frac{tr\left|M^{-1}\left(\xi_A^*\right)\right|}{tr\left|M^{-1}\left(\xi\right)\right|}$$ Where ξ_A^* is A optimal

The optimum design for A is 1.862176, making this design (0.478) 48% efficient. Also the Ge that is the G efficiency is available as the standard of design quality is estimated to (0.575) 56% efficient

D-optimality Criterion: Is the most popular design criterion in the life applications, which introduced by Wald (1943), put the emphasis on the quality of the parameter estimates. D-optimality criterion is also known as the determinant criterion the aim of D-optimality is essentially a parameter estimation criterion. This was called lately, D–optimality by Kiefer and Wolfowitz (1959). The D-optimality is define as $\max_{x_i i=1,\dots n}\left|X^{1}X\right| \equiv \min_{x_i i=1,\dots n}\left|\left(X^{1}X\right)^{-1}\right|$ which means maximizing the determinant of the information matrix, or equivalently, minimizing the determinant of the inverse of the information matrix. D-efficiency of a design ξ is defined as:

$$D(\xi)=\left[\frac{|M(\xi)|}{M(\xi_D^*)}\right]^{1/p}$$ Where ξ_D^* is D optimal

The optimum design for D is 0.8513693, making this design (0.944) 94% efficient. Also the Ge that is the G efficiency is available as the standard of design quality is estimated to (0.945) 95% efficient

4. Discussion

Based on the data simulated from R statistical package, Poisson regression model were employed according to McCullagh and Nelder (1989), it is the simplest way of modeling count data, quasi Poisson model were used to check for over dispersion in data and it was confirm by the dispersion parameter greater than 1 but given the same regressor as Poisson regression model. Another formal way of modeling over dispersion in Poisson regression is the use of negative binomial regression model given the model estimates dispersion parameter as 158.6768; setting theta = 159 when generating random variates

Furthermore, the optimal design were also employed using the optimality criterion that is the A and D optimality criterion and find which design form best among the two (2) optimality criterion using the design efficiency (D-efficiency) and global efficiency (G-efficiency) respectively. The design with highest D and G efficiency shows the best, thus the D-optimality criterion has the highest efficiency.

5. Conclusions

From the discussion of result and analysis above the following conclusion were made

Generalized linear model

- The Poisson regression model was used to fit a model, the parameters of the fitted model were found to be significant.
- Quasi Poisson regression was use to test for over dispersion and it was found that there is over dispersion in Poisson regression model which lead to use of negative binomial regression model
- Using the AIC the Poisson regression model give an appropriate model having the minimum AIC in the analysis

Optimal design

From the difference optimality criteria in optimal design that is A- and D-optimality use for the analysis of the data, design efficiency was used to compare the two designs.

- The Design efficiency for D-optimality was obtain to be 0.944 (94%) and the A-optimality to be 0.478 (48%) respectively which implies that the two criteria are efficient but the D-optimality is more efficient than the A-optimality.
- The D-optimality with the highest Design efficiency shows the best design from other optimality criteria

Appendices

R statistical command use for the analysis of data

set.seed(1234)# is use to set a seed and get the same and accurate result when repeating many number of time for the analysis

```
library(MASS)# Package for negative binomial model
beta0=1
beta1=0.2
beta2=0.4
x1=runif(300,0,1.5)
x2=runif(300,0,2.5)
mu=exp(beta0+beta1*x1+beta2*x2)
y=rpois(300,mu)
dat=data.frame(y,x1,x2)
fit1=glm(y~x1+x2,family=poisson,data=dat)
summary(fit1)
fit2=glm(y~x1+x2,family=quasipoisson,data=dat)
summary(fit2)
fit3=glm.nb(y~x1+x2,data=dat)# Negative binomial model
summary(fit3)
# D-optimality criterion
library(AlgDesign)# Package for optimal design
desD=optFederov(y~x1+x2,data=dat,eval=Tcrit=D)
desD
# A-optimality criterion
desA=optFederov(y~x1+x2,data=dat,eval=T,crit="A")
desA
```

References

Agresti, A., 2002. Categorical Data Analysis Wiley- Interscience, New York, USA.

Ash, A., Hedayat, A., 1978. An Introduction to Design Optimality with an Overview Literature. Communicat. Statist., 14, 1295-1325.

Atkinson, A.C., Chaloner, K., Herzberg, A.M., Juritz, J., 1993. Optimum Experimental Designs for Properties of a Comp. Mod. Biometr., 49, 325-337.

Atkinson, A.C., Donev, A.N., 1992. Optimum Experimental Designs, Clarendon Press, Oxford.

Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum Experimental Designs, with SAS. Oxford Univ. Press.

Box, G.E.P., Hunter, W.G., Hunter, S.J., 1978. Statistics for Experimenters, John Wiley & Sons, Inc., New York, NY.
Chernoff, H., 1953. Locally Optimal Designs for Estimating Parameters. Annal. Mathemat. Statist., 24, 586-602.
Fedorov, V.V., 1972. Theory of Optimal Experiments, Academic Press, New York.
Khuri, A.I., Mukherjee, B., Sinha, B.K., Ghosh, M., 2006. Design Issue for Generalized Linear Model. A Rev. Statist. Sci., 21,376-399.
Kiefer, J., 1959. Optimum Experimental Design (With Discussion). J. Roy. Statist. Soc. Ser. B., 21, 272-319.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models 2nd Edition. London. Chapman Hall.
Montgomery, D.C., 2000. Design and Analysis of Experiments, Fifth Edition. John Wiley Sons, New York, NY.
Myers, R.H., Montgomery, D.C., Vining, G.G., 2002. Generalized Linear Models. John Wiley Sons, New York.
Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. Roy. Statist. Soc., Series A 135, 370-384.
R Development Core Team., 2012. R: A language and environment for statistical computing. R Foundat. Statist. Comput., Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
Smith, K., 1918. on the standard deviation of adjusted and interpolated values of an observed polynomial Function and its Constants and the Guidance they give toward a proper cho. Distribut. Observat. Biometr., 12, 1-85
Wald, A., 1943. On the Efficient Design of Statistical Investigation. Mathemat. Statist., 14, 134-140.
Wheeler, R.E., 2004. optFederov. AlgDesign. The R project for statistical computing http://www.r-project.org/.
Yang, J., Mandal, A., Majumdar, D., 2012. Optimal designs for two-level factorial experiments with binary response. Statist. Sin., 22, 885–907.