Open Journal of Statistics, **2013, 3, 293-298**
http://dx.doi.org/10.4236/ojs.2013.35034 Published Online October 2013 (http://www.scirp.org/journal/ojs)

Generalized Estimating Equations for Repeated Measures Logistic Regression in Mosquito Dose-Response

Gabriel Otieno[1], Gichihu A. Waititu[1], Daisy Salifu[2]
[1]Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[2]African Insect Science for Food and Health (ICIPE), Nairobi, Kenya
Email: gabriel2ke@gmail.com, agwaititu@yahoo.com, salifudp@yahoo.com

Received June 26, 2013; revised July 26, 2013; accepted August 4, 2013

ABSTRACT

Dose-response studies in arthropod research usually involve observing and collecting successive information at different times on the same group of insects exposed to different concentrations of stimulus. When the same measure is collected repeatedly over time, the data become correlated and Probit Analysis technique which is the standard method in analyzing bioassay experiments data cannot be used. Lethal time is estimated when the speed of kill is of interest since mortality varies over time. We evaluate a complementary approach, repeated measures logistic regression using Generalized Estimating Equations (GEE), for lethal time determination in mosquito dose response. Mortality data from anopheles larva exposed to 3 botanical extracts (B,C,E) at 4 concentration levels: 500 mg/ml, 250 mg/ml, 50 mg/ml and 12.5 mg/ml were used. The result shows the estimated LT_{50} values with concentration 500 mg/ml being the most virulent chemical for extract B (LT_{50} = 10.3 hrs), C (LT_{50} = 7.2 hrs) and E (LT_{50} = 10.3 hrs). The least virulent chemical was concentration 12.5 mg/ml for extract B (LT_{50} = 52.1 hrs), C (LT_{50} = 70.7 hrs) and E (LT_{50} = 55.0 hrs). We conclude that repeated measures of logistic regression via GEE can be used as a tool to estimate LT_{50} more effectively in repeated measures of arthropod data.

Keywords: Dose-Response; GEE; Lethal Time; Probit Analysis; Repeated Measures

1. Introduction

Dose-response studies in arthropod research usually involve observing and taking successive measurements of insects' mortality on groups of insects subjected to different concentrations of stimulus [1,2], giving rise to repeated measures data. Mortality data collected several times on the same group of organisms at several concentrations over time are usually correlated [1-4] and cannot be analyzed using standard Probit analysis technique [5,6] which is the usual way of analyzing data from bioassay experiment [1,2]. Probit analysis is adequate if the responses are independent, true for data collected at once after a given time point [6]. In arthropod dose response studies, samples of insets are usually exposed to several concentrations of insecticide to determine the concentration that will kill 50% (LT_{50}) of the insects within a given time span [6,7]. Effect of time on the percentage of kill at one or several concentration is of importance when the interest is in the speed of kill because mortality varies with time [1,2].

Given the correlated measurements in dose-response studies and when the interest is in the speed of kill, one has to move on to alternative method which accounts for the correlation in the data while estimating lethal time and of such methods is the Generalized Estimating Equations (GEE) [8]. With GEE correlated data can be modeled with output that looks similar to generalized linear models (GLMs) with independent observations by accounting for the within-subject covariance structure [9,10]. The available covariance structures specify how observations within a subject or cluster are correlated with each other [11].

Arthropod dose-response data may have a binary repeated measures response and therefore GEE in a logistic regression setting will be a good way to model the data [8,12,13]. Usually Logistic regression is a Generalized Linear Model (GLM) method for analyzing binary outcome [14,15] but ignores the correlated nature of the data. In this paper the use of repeated measures logistic regression using GEE is considered as complementary approach to LT_{50} estimation to address the limitation of Probit Analysis in estimating LT_{50} for correlated mosquito dose re-

sponse data. GEE for repeated measures logistic regression was used because the data were binary and correlation because time was to be taken into account.

2. Methods

The data used in this paper were from a laboratory experiment on the effect of botanical extracts on mortality of larvae of anopheles mosquito (*Anopheles gambiae*) as part of malaria control project. Several botanical products were studied but in this paper we chose only three botanicals namely B,C,E and control D. The botanicals were studied at four concentration levels: 12.5 mg/ml, 50 mg/ml, 250 mg/ml and 500 mg/ml. Fifty larvae were dipped in glass beaker containing the specific botanical products at a specific concentration. Each concentration with specific botanical extracts was replicated three times. The response variable was larval mortality observed at 12 hrs, 24 hrs, 36 hrs, 48 hrs, 60 hrs and 72 hrs after exposure. There was no death in control which consisted of water only and hence does not appear in the analysis. The data collected had three factors; botanical extracts, concentrations and time. The data set was created for each extract at each concentration level as shown in Appendix 1. GEE model in a logistic regression setting was used to estimate LT_{50}. R statistical software version R 2.14.1 was used in the data analysis.

2.1. Logistic Regression

Logistic (logit) regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables [14, 15]. In arthropod dose-response mortality data is a set of Bernoulli trials which is a special case of Binomial distribution. The values of response Y_i (mortality status) are 1 if there is a success and 0 otherwise. The binary response is the mortality status of 50 mosquito larva at time 12 hrs, 24 hrs, 36 hrs, 48 hrs, 60 hrs and 72 hours at a given concentration level. Generalized linear models (GLM) are a generalization of standard linear regression so that the response variables may have a distribution other than the Gaussian [14,15]. Logistic regression is the appropriate GLM when the data follows Bernoulli or Binomial distribution.

For a binary response variable Y (mortality status), and a set of 1 predictor variable (time), X_1 at a given concentration level with a logistic transformation or logit function, the logistic regression will be given by

$$\operatorname{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \varepsilon \qquad (1)$$

where π is the probability of success, β_0 is the intercept, β_1 is the regression coefficient for each corresponding predictor variable, X_1 (time), at a given concentration and ε is the error of the prediction [14,15].

2.2. Expressing Lethal Time (LT50) Using Logistic Regression

Consider Equation (1) in the form

$$\operatorname{logit}\left(\pi(\text{time})\right) = \beta_0 + \beta_1(\text{time}) \qquad (2)$$

LT_{50} is the time at which $\pi(\text{time})$ equals 0.5. [3,6] and by substituting $\pi(\text{time})$ with 0.5 in Equation (1) gets

$$LT_{50} = -\frac{\beta_0}{\beta_1} \qquad (3)$$

Any tests comparing lethal time values should include confidence limits of the estimated statistics [1]. Based on the asymptotic approximation, the variance of the LT_{50} computed using the delta method [4,12] is

$$\begin{aligned}\operatorname{Var}(LT_{50}) = {} & \frac{1}{\beta^2}\operatorname{Var}(\beta_0) + \frac{\beta_0^{\,2}}{\beta^4}\operatorname{Var}(\beta) \\ & + 2\cdot\frac{1}{\beta^2}\cdot\frac{\beta_0^{\,2}}{\beta^4}\cdot\operatorname{Cov}(\beta_0,\beta)\end{aligned} \qquad (4)$$

and hence an approximate 95% confidence interval (CI) for the LT_{50} is given by

$$\left[LT_{50} - 1.96\cdot\sqrt{\operatorname{Var}(LT_{50})},\ LT_{50} + 1.96\cdot\sqrt{\operatorname{Var}(LT_{50})}\right] \qquad (5)$$

To account for correlation effect due to time (repeated measures), GEE is used to estimate the parameters β_0 and β_1 by specifying the correlation structure [10,16,17] to permit for the calculation of robust estimates for the standard error of the regression coefficients

2.3. Generalized Estimating Equations

Let Y_{ij}, $i = 1,\cdots,n$, $j = 1,\cdots,n_i$ denote the mortality status of mosquito larva j after exposure to at a given concentration i for a given botanical extract ($Y_{ij} = 1$, dead and $Y_{ij} = 0$, alive). Let X_{ij} be the time taken by mosquito larva j to die after being exposed to concentration i. Y_{ij} is assumed to follow a Bernoulli distribution when the probability that mosquito larva is dead and is denoted by μ_{ij}, that is $\mu_{ij} = P(Y_{ij})$ and this is also equal to the expected death $E(Y_{ij}) = \mu_{ij}$.

The marginal logistic regression model for the data is

$$\begin{aligned}&\operatorname{logit}(\mu_{ij}) = \beta_0 + \beta X_{ij} \\ &\operatorname{Var}(Y_{ij}) = \mu_{ij}(1-\mu_{ij}) \\ &\operatorname{Corr}(Y_{ij}, Y_{ik}) = \alpha_{jk}\end{aligned} \qquad (6)$$

In this model the number of observations per cluster (time intervals) is small and in a balanced and complete design, hence unstructured correlation matrix [9,11,18].

The observations are correlated with no assumptions of the structure.

To use GEE in estimating, there are three-part specification; the conditional expectation of each response, the conditional variance of each Y_{ij} given the covariates and the covariance (correlation) matrix [10,16,17].

Let the marginal regression model to be:

$$g\left(E\left[Y_{ij}/X_{ij}\right]\right)=X_{ij}'\beta \tag{7}$$

where X_{ij} is a $p\times 1$ vector of covariates, β consists of the p regression parameters of interest (time) $g(.)$ is the link function, and Y_{ij} denotes the j^{th} outcome $\left(\text{for } j=1,\cdots,J\right)$ for the i^{th} mosquito larva/subject $\left(\text{for } j=1,\cdots,J\right)$. For this paper the link function chosen was the logit link for binary data [9].

The GEE equation for vector β or the regression model (score) is given by

$$g\left(\beta\right)=\sum_{i=1}^{k}D_i^{T}V_i^{-1}\left(Y_i-\mu_i\right)=0 \tag{8}$$

where D_i is the matrix of derivatives $\dfrac{\partial\mu_i}{\partial\beta_j}$, V_i is the "working" covariance matrix of Y_i.

Let $R_i\left(\alpha\right)$ be an $n_i\times n_i$ "working" correlation matrix that is fully specified by the vectors of parameter α [8,9,16,18]. The variance-covariance matrix, part of the model used in the estimating equation, is:

$$V_i=\phi A_i^{\frac{1}{2}}R_i\left(\alpha\right)A_i^{\frac{1}{2}} \tag{9}$$

where ϕ is a glm dispersion parameter to allow for over dispersion, A_i is a diagonal matrix of variance functions $\left(\text{Var}\left(Y_{ij}\right)\right)$ *i.e.* $\text{Var}\left(Y_{ij}\right)=\mu_{ij}\left(1-\mu_{ij}\right)$, and $R_i\left(\alpha\right)$ is the correlation matrix of Y_i.

2.4. Working Correlation Matrix

The working correlation matrix is usually unknown and must be estimated. It is estimated in the iterative fitting process by using the current value of the parameter vector β. Common choices for the correlation structure within GEE include Independent, exchangeable, autoregressive (AR(1)), unstructured, M-dependent and User fixed [8-10,16-18].

2.5. Choosing the Correlation Structure in GEE

Quasi-likelihood Information Criterion (QIC) is usually applied to models fir by GEE to find an acceptable working correlation structure giving the least QIC [11].

$$QIC=-2Q\left(\hat{\mu};I\right)+2\text{trace}\left(A_I^{-1}V_R\right) \tag{10}$$

where I is the independent covariance structure used to calculate the quasi-likelihood. $\hat{\mu}=g^{-1}\left(X\hat{\beta}\right)$ and $g^{-1}(\ )$ is the inverse link function for the model (logit). A_I^{-1} is the variance matrix under the assumption of independence model and V_R is the robust variance estimator obtained from a general working covariance structure R. Prior knowledge on how the data was collected may also guide in choosing the best correlation structure to reflect the manner in which the data was collected [11]. LT_{50} is then estimated using repeated measures logistic regression which uses GEE as an implementing tool.

Given a mean model, μ_{ij}, and variance structure, V_i, ("working" covariance matrix of Y_i), the parameter estimates will be given by solving $g\left(\beta\right)=0$ which is usually obtained via the Newton-Raphson algorithm or via iterations [16,17].

The covariance of matrix $\beta,\left(\text{Cov}\left(\beta\right)\right)$, is estimated using the model-based estimator and the empirical or robust estimator [9,10,17].

The model-based estimator of the covariance matrix of β is given by

$$\text{Cov}\left(\beta\right)_m=\sum\nolimits_m\left(\beta\right)=I_0^{-1} \tag{11}$$

where

$$I_0=\sum_i^k\frac{\partial\mu_i'}{\partial\beta}V_i^{-1}\frac{\partial\mu_i}{\partial\beta}=D^TV^{-1}D \tag{12}$$

for this case $\text{Cov}\left(\beta\right)_m$ consistently estimates $Cov\left(\beta\right)$ if the mean model and the working correlation are correct [8,9].

The empirical or robust estimator of the covariance matrix of β is given by

$$\text{Cov}\left(\beta\right)e=\sum\nolimits_e\left(\beta\right)=I_0^{-1}I_1I_0^{-1} \tag{13}$$

where

$$\begin{aligned}I_1&=\sum_{i=1}^{k}\frac{\partial\mu_i}{\partial\beta}V_i^{-1}\text{Cov}\left(Y_i\right)V_i^{-1}\frac{\partial\mu_i}{\partial\beta}\\&=D^TV^{-1}\left(Y-\mu\right)\left(Y-\mu\right)^TV^{-1}D\end{aligned} \tag{14}$$

for this case $\text{Cov}\left(\beta\right)_e$ is a consistent estimator of $Cov\left(\beta\right)$ even if the working correlation is misspecified [8,9].

3. Results

The results of lethal time determination for mosquito dose-response using repeated measures logistic regression via GEE are presented in **Table 1**. Across the three extracts concentration 500 mg/ml was the most potent chemical, followed by concentration 250 mg/ml, concentration 250 mg/ml and concentration 50 mg/ml in that order (**Table 1**, **Figures 1** and **2**).

Botanical extracts B, C and E were significantly different from each other in terms of insect mortality across

Table 1. LT_{50} estimates from repeated measures logistic regression using GEE.

Extract	Concentration (mg/ml)	LT_{50} (hrs)	95% CI for LT_{50} (hrs)
B	12.5	52.1	50.5 - 53.7
B	50	23.0	16.8 - 29.3
B	250	12.3	7.4 - 17.2
B	500	10.3	1.51 - 19.1
C	12.5	70.7	69.3 - 72.0
C	50	43.4	42.0 - 44.7
C	250	21.5	19.1 - 23.9
C	500	7.2	4.3 - 10.1
E	12.5	55.0	52.6 - 57.3
E	50	16.6	11.57 - 21.7
E	250	12.2	6.16 - 18.2
E	500	10.3	9.5 - 11.3

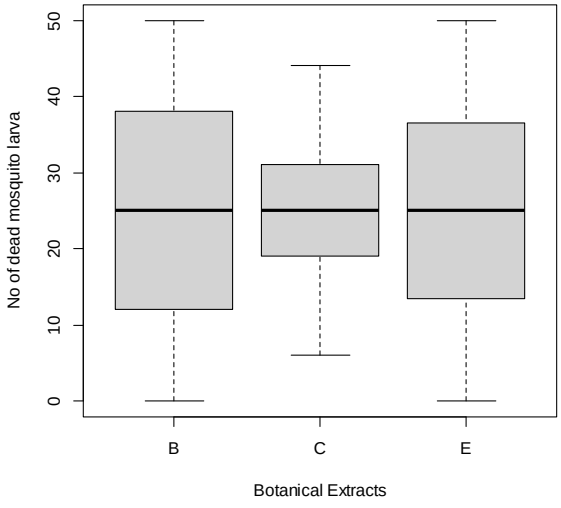

Figure 1. Box plot for extracts B, C and E.

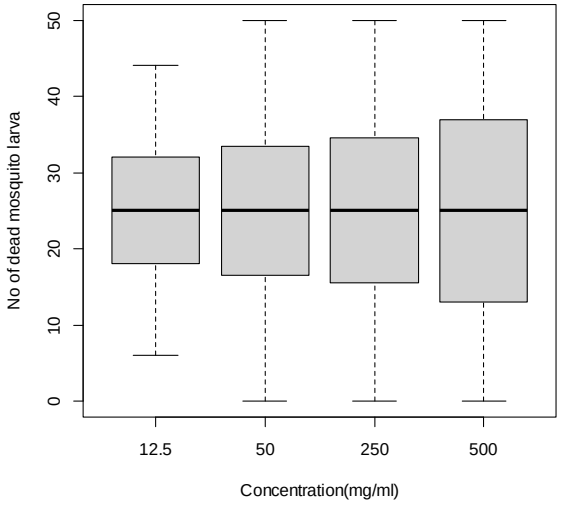

Figure 2. Box plot for concentration 12.5 mg/ml, 50 mg/ml, 250 mg/ml and 500 mg/ml.

all time points (**Figure 1**). The concentration levels 12.5 mg/ml, 50 mg/ml, 250 mg/ml and 500 mg/ml were different from each other in terms of insect mortality across all the time intervals (**Figure 2**).

Estimates of the lethal time (LT_{50}) values with 95% CI for the different concentrations for the different botanical extracts against anopheles mosquito are shown in **Table 1**. The LT_{50} values ranged between 10.3 hrs to 52.1 hrs for extract B; between 7.2 hrs to 70.7 hrs for extract C and between 10.3 hrs to 55 hrs for extract E. The LT_{50} values for the different concentration levels ranged between 52.1 hrs to 70.7 hrs for concentration 12.5 mg/ml; between 16.6 hrs to 43.4 hrs for concentration 50 mg/ml; between 12.2 hrs to 21.5 hrs for concentration 250 mg/ml; and between 7.2 hrs to 10.3 hrs for concentration 500 mg/ml.

4. Discussion

This paper has used repeated measures logistic regression using GEE method to estimate LT_{50} in repeated measures for mosquito (arthropod) dose-response.

Estimating LT_{50} is of importance when the interest is in the speed of kill since mortality varies with time. It's also of importance because observations made on the same group of organisms at different times are correlated and hence standard Probit analysis will not be applicable [1,18]. Repeated measure logistic regression using GEE was able to estimate LT_{50} for the different concentration levels together with their corresponding confidence intervals [1,4,12,13,19].

The analysis showed that concentration 500 mg/ml was the most potent chemical while concentration 12.5 mg/ml was least potent chemical. In studying the lethal effects of concentrations on mortality, higher concentration levels are usually more effective in regards to mortality [2,12,13,19] which seems to have been reflected in the estimated LT_{50} for the different concentrations. Concentration 500 mg/ml was the most potent chemical since it took shorter time to kill half of the insects' population. Further research should be done to ascertain the claim of the estimated LT_{50} to rule out if there are effects of some other factors. LT_{50} and the confidence intervals of the estimates in this paper were similar with results from the same methods but applied in a different setting [12,19] to show that the method was versatile for analyzing repeated measures dose response data from arthropod studies.

The exact time of kill was not known in the GEE approach since time was used cumulatively to estimate if the mosquito larva has been killed at a particular time point. Effective data collection methods and use of existing methods of estimating LT_{50} should be used in a complementary fashion. Unstructured correlation matrix was the only one used in repeated measures logistic re-

gression via GEE. Wider comparisons and use of QIC should be considered to make the research more representative.

The combined use of GEE approach together with other existing analytical methods for bioassay data may improve the way how repeated measures arthropod dose-response data is being analyzed when the speed of kill is of interest [1,18].

As a complementary approach to Probit analysis and other existing methods for analyzing data from bioassay experiments, repeated measures logistic regression via GEE can be used as a tool to estimate LT_{50} more effectively in repeated measures arthropod data. Wider exploration of GEE techniques and further testing and refinement are needed to fully develop its promising capabilities.

5. Acknowledgements

The authors acknowledge Jomo Kenyatta University of Science and Technology (JKUAT), African Insect Science for Food and Health (ICIPE) and Regional Universities Forum for Capacity Building in Agriculture (RUFORUM) for their support. This work was undertaken by the lead author as an MSc research at JKUAT.

REFERENCES

[1] J. E. Thorne, D. K. Weaver, V. Chew and J. E. Baker, "Probit Analysis for Correlated Data: Multiple Observations over Time at One Concentration," *Journal of Economic Entomology*, Vol. 88, No. 5, 1995, pp. 1510-1512.

[2] J. L. Robertson and H. K. Preisler, "Pesticide Bioassays with Arthropods," CRC, Boca Ratn, 1992.

[3] H. K. Preisler and J. L. Robertson, "Analysis of Time-Dose-Mortality Data," *Journal of Economic Entomology*, Vol. 82, No. 6, 1989, pp. 1534-1542.

[4] L. Thomsen and J. Eilenberg, "Time-Concentration Mortality of Pieris Brassicae (Lepidoptera: Pieridae) and Agrotis Segetum (Lepidoptera: Noctuidae) Larvae from Different Destruxins," *Entomological Society of America*, Vol. 29, No. 5, 2000, pp. 1041-1047.

[5] D. J. Finney, "Statistical Methods in Biological Assay," 2nd Edition, Griffine, London, 1964.

[6] D. J. Finney, "Probiy Analysis," Cambridge University Press, Cambridge, 1971.

[7] M. Eaton and S. A. Kells, "Use of Vapor Pressure Deficit to Predict Humidity and Temperature Effects on the Mortality of Mold Mites, Tyrophagus Putrescentiae," *Experimental and Applied Acarology*, Vol. 47, No. 3, 2009, pp. 201-213. http://dx.doi.org/10.1007/s10493-008-9206-2

[8] M. E. Stokes, C. S. Davis and G. G. Koch, "Categorical Data Analysis Using the SAS System," SAS Institute, Inc., Cary, 2000.

[9] A. Ziegler, C. Kastner and M. Blettner, "The Generalised Estimating Equations; An Annotated Bibliography," *Biometrical Journal*, Vol. 40, No. 2, 1998, pp. 115-139. http://dx.doi.org/10.1002/(SICI)1521-4036(199806)40:2<115::AID-BIMJ115>3.0.CO;2-6

[10] S. L. Zeger and K. Y. Liang, "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, Vol. 42, No. 1, 1986, pp. 121-130. http://dx.doi.org/10.2307/2531248

[11] J. W. Hardin and J. M. Hilbe, "Generalized Estimating Equations," Chapman and Hall, New York, 2003.

[12] D. M. Bugeme, H. I. B. Knap, A. K. Wanjoya and N. K. Maniani, "Influence of Temperature on Virulenec of Fungal Isolates of Matarhizuim Anissopliae and Beaureria Bassania to the Two-Spotted Spider Mite Tetrancychus Uriticae," *Mycopathologia*, Vol. 167, No. 4, 2009, pp. 221-227. http://dx.doi.org/10.1007/s11046-008-9164-6

[13] M. Latifian and B. Rad, "Pathogenicity of Entomopathogenic Fungi Beaureria Bessiana (Balsamo) Vuillmin, Beaveria Brongniartii Saccardo and Metarhizium Anisopliae Metsch to Adult Oryctes Elegans Prell and Effects on Feeding and Fundicity," *The International Journal of Agriculture Sciences*, Vol. 4, No. 14, 2012, pp. 1026-1032.

[14] P. McCullag and J. A. Nelder, "Generalized Linear Models," Chapman and Hall, London, 1983.

[15] P. McCullagh and J. A. Nelder, "Generalized Linear Models," 2nd Edition, Chapman and Hall, London, 1989.

[16] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev and G. M. Smith, "Mixed Effects Models and Extensions in R," Sringer Science and Business Media, New York, 2009. http://dx.doi.org/10.1007/978-0-387-87458-6

[17] K. Y. Liang and S. L. Zeger, "Longitudina Data Analysis Using Generalized Linear Models," *Biometrika*, Vol. 73, No. 1, 1986, pp. 13-22. http://dx.doi.org/10.1093/biomet/73.1.13

[18] M. Crowder, "On the Use of a Working Correlation Matrix in Using Generalized Linear Models for Repeated Measures," *Biometrika*, Vol. 82, No. 2, 1995, pp. 407-410. http://dx.doi.org/10.1093/biomet/82.2.407

[19] D. M. Mburu, L. Ochola, N. K. Maniani, P. G. N. Njagi, L. M. Gitonga, M. W. Ndungu, A. K. Wanjoya and A. Hassanali, "Relationship between Virulence and Repellency of Entomopathogenic Isolates of Metarhizium Anisopliae and Beaureria Bassiana to the Termite Macrotermes Michaelseni," *Journal of Insect Physiology*, Vol. 55, No. 9, 2009, pp. 774-780. http://dx.doi.org/10.1016/j.jinsphys.2009.04.015

Appendix

```
>geedata
   extract time dose rep total success prop IDD
1        E   12  12.5   1   50     8    0.16   1
2        E   12  12.5   2   50    10    0.20   1
3        E   12  12.5   3   50    12    0.24   1
4        E   24  12.5   1   50    12    0.24   2
5        E   24  12.5   2   50    15    0.30   2
6        E   24  12.5   3   50    13    0.26   2
7        E   36  12.5   1   50    16    0.32   3
8        E   36  12.5   2   50    15    0.30   3
9        E   36  12.5   3   50    18    0.36   3
10       E   48  12.5   1   50    25    0.50   4
11       E   48  12.5   2   50    20    0.40   4
12       E   48  12.5   3   50    23    0.46   4
13       E   60  12.5   1   50    30    0.60   5
14       E   60  12.5   2   50    28    0.56   5
15       E   60  12.5   3   50    26    0.52   5
16       E   72  12.5   1   50    33    0.66   6
```

Appendix 1. Data format for repeated measure logistic regression using GEE.

