



Analysis of saturated 2-level factorial experiments with gamma distributed response variable

Sangita Birajdar (SC Patil)

MIT ACS College, Alandi (D), Pune, Maharashtra, India

Abstract

This paper assesses and compares the performance of Generalized linear models (GLM) with respect to the traditional transformation approach for saturated 2-level factorial experiments for Gamma distributed response variable. The evaluation is based on the coverage and expected length of Confidence Interval (CI) for the expected response for factorial effects under these approaches at the prescribed confidence coefficient. The traditional transformation approach outperforms GLM approach for its non-canonical link for two replications with respect to coverage probabilities. However the extent of out-performance does not attain the desired confidence level. An ad-hoc modification to the CI for transformation approach and non-canonical link of GLM approach is suggested, that shows decisive improvement over their performances for small shape parameters.

Keywords: confidence intervals, factorial experiment, generalized linear models, log transformation

1. Introduction

The ANOVA based F-test is a widely used technique for the analysis of designed experiments. The optimality properties of this test heavily depend on normality and homogeneity of variances of the response variable. But many real life investigations involve non-normal response variables implying non-homogeneous variances and non-linear relationship with the input factors. Moreover, a small number of replications is a common practice in many investigations. Lewis et.al. (2001a) gave many examples of designed experiments with non-normal responses for which a common practice is to use a suitable transformation on the response variable to achieve conditions close to those of the normal scenario. For gamma distributed response variable log transformation is commonly used variance stabilizing transformation which is refereed as 'LOG approach' in this work.

Several authors have used GLM, introduced by Nelder and Wedderburn (1972) as a powerful alternative to data transformation. Myers and Montgomery (1997) observed that, models built using GLM perform better than those built with the traditional approaches with respect to response estimation and prediction when applied on some real data sets. Lewis et.al. (2001b) attempted a simulation study on analysis of designed experiments involving non-normal response variables using GLM and advocated the use of GLM for small samples. Patil and Kulkarni (2011) attempted an in-depth comparative study among the traditional approaches (ANOVA, log transformation) and GLM approach for its non-canonical link for 2ⁿ full factorial experiments with an Exponential response variable. The comparison was based on the coverage and expected length of confidence intervals (E(LOCI)) for the expected response on theoretical background and simulation study. They strongly suggested not to use classical ANOVA approach as it is unstable and misleading. In this sequel the present work evaluates the

performance of GLM approach with respect to traditional transformation approach for small sample sizes that could be typically encounter in industrial experiments for saturated 2-level factorial experiments with Gamma distributed response variable when the shape parameter is known. For a gamma distributed response, the non-canonical link selects the 'Log link' which is refereed as 'GLM-LOG' and the canonical link selects the 'Reciprocal link' which is refereed as 'GLM-REC' throughout this work. Due to some drawbacks of canonical link of GLM, for example, negative values of the estimated response may be encountered, Patil and Kulkarni (2011) chosen only non-canonical link for exponentially distributed response variable. For Gamma distributed response variable these drawbacks are not encountered when the shape parameter is greater equal 2.

The family of Gamma distribution is popularly used for modeling the lifetime response variables. In many industrial designed experiments, the response variable is a lifetime random variable. Lifetime of produced items, Annual claims received by an insurance company, Survival time of patients having certain illness etc. are some examples of Gamma distributed response variables. Performance of the above discussed methods in terms of E(LOCI) and corresponding coverages (or lower bounds (L.B.) to the coverages) of the interval estimates of expected response is investigated for Gamma distributed response variable in Section 2. Theoretical results are not directly comparable, which suggest simulation based comparison. Section 3 reports the simulation based comparison of the performance of the above discussed methods in terms of E(LOCI) and corresponding coverages for expected response. Section 4, suggests a modification to GLM approach for its non-canonical link and modification to transformation approach based CI for small values of shape parameter. Section 5 gives recommendations and concluding remarks.

2 Theoretical Details

2.1 Setup, Underlying Model and CI for Three Approaches

The set up considered here is that of the 2-level full saturated factorial designs with $r \geq 1$ replications. Let I_k denote an identity matrix of order k . The underlying model can be viewed as a regression model with design matrix (cf. Myers and Montgomery (1997)),

$$X_{n \times k} = (A, \dots, A) \quad (1)$$

where $n = rk$ and A is a $k \times k$ symmetric Hadamard matrix corresponding to a single replicate of the experiment. Throughout this work, it is assumed that, the response variable Y is a Gamma distributed random variable with a known shape parameter $\nu \geq 2$. $\underline{Y}$ denotes the vector of all observations. $\hat{\underline{Y}}$ denotes the estimated response, μ_i denotes the expected response corresponding to the i^{th} regressor vector X_i , the i^{th} row of the coefficient matrix X . The components of the vector of parameter $\underline{\beta}$ correspond to the intercept term and various factorial effects.

$Y_{ij} \sim \text{Gamma}(\nu, \lambda_i)$, $i = 1, 2, \dots, k$, $j = 1, 2, \dots, r$ are observations on Y . $\bar{Y}_{i.} = \frac{1}{r} \sum_{j=1}^r Y_{ij}$, $W_{ij} = \log(Y_{ij})$ is the log

transformed response, $S_y^2 = (1/r) \sum_{i=1}^k S_i^2$ where,

$$S_i^2 = (r-1)^{-1} \sum_{j=1}^r (Y_{ij} - \bar{Y}_{i.})^2 \text{ and } \bar{W}_{i.} \text{ and } S_w^2 \text{ are similarly}$$

defined. Note that, the LOG approach assumes a constant variance σ_w^2 for all components of $\underline{W}$. $t_{(n-k), \alpha/2}$ and Z_α

are $(1 - \alpha)^{th}$ quantiles of the t-distribution with $n - k$ degrees of freedom (d.f.) and the standard normal distribution respectively. For illustration, the confidence coefficient $1 - \alpha = 0.95$ is used throughout the work.

For GLM approach, the model is $E \underline{Y} = \underline{\mu} = g^{-1}(\underline{\eta})$;

$\underline{\eta} = X' \underline{\beta} = g(\underline{\mu})$ where $g(\cdot)$ is an appropriate monotonic and differentiable link function. The distribution of the response variable $\underline{Y}$ is assumed to be a member of the exponential family. For more theoretical details of generalized linear models refer to McCullagh and Nelder (1989). For a gamma distributed response $\underline{Y}$, the non-canonical link selects the log link $\eta_i = \log(\mu_i)$ and for this link, the underlying model is

$$E(Y_{ij}) = \mu_i = \lambda_i \nu = \exp(X_i' \underline{\beta}) \quad i = 1, \dots, k, j = 1, \dots, r. \quad (2)$$

While for canonical link selects the reciprocal link $\eta_i = (\mu_i)^{-1}$ and the underlying model is

$$E(Y_{ij}) = \mu_i = \lambda_i \nu = (X_i' \underline{\beta})^{-1} \quad i = 1, \dots, k, j = 1, \dots, r. \quad (3)$$

where X_i' is the i^{th} row of A in the design matrix (1) and $\underline{\mu} = (\mu_1, \dots, \mu_k)$ is the mean response vector corresponding to a single replicate. Parameter estimation in the GLM is

performed using the method of maximum likelihood estimation. Under the full model, the estimated response in a

closed form; $\hat{Y}_{ij} = \hat{\mu} = \bar{Y}_{i.}$. Consequently $\hat{\underline{\beta}}$ is also available in a closed form which is not the case in the non-full models. The CI for mean response $\mu_i = \lambda_i \nu$ for GLM approach for its canonical and non-canonical link is given by, (cf. Myers and Montgomery (1997)).

2.2 Theoretical E(LOCI) for Three Approaches

Table 1: Confidence Interval for GLM Approach

Method	CI for mean Response μ_i	$\hat{\sigma}_i$
GLM- LOG	$\exp(X_i' \hat{\underline{\beta}} \pm Z_{\alpha/2} \hat{\sigma}_i)$	$\hat{\sigma}_i = \{(a(\phi))^2 x_i' (X' \hat{V} \Delta X)^{-1} x_i\}^{1/2}$
GLM- REC	$(X_i' \hat{\underline{\beta}} \pm Z_{\alpha/2} \hat{\sigma}_i)^{-1}$	$\hat{\sigma}_i = \{(a(\phi))^2 x_i' (X' \hat{V} X)^{-1} x_i\}^{1/2}$

Here $\hat{V}$ is the diagonal matrix whose i^{th} diagonal element is an estimate of $\text{Var}(Y_i)$ and Δ is the diagonal matrix whose i^{th} diagonal element Δ_i is the derivative of $(-1/\mu_i)$ with respect to $X_i' \underline{\beta}$. For the design matrix (1), noting that, for gamma distributed response variable $a(\phi) = \nu^{-1}$, $\Delta_i = \exp(-X_i' \underline{\beta})$ and $\text{Var}(Y_i) = \nu^{-1} \mu^2$.

This simplifies to $\hat{\sigma}_i = 1/\sqrt{r\nu}$ for GLM-LOG approach and $\hat{\sigma}_i = (X_i' \underline{\beta} \sqrt{r\nu})^{-1} = \bar{Y}_{i.}/\sqrt{r\nu}$ for GLM-REC approach. A further simplification leads to the CI and expected length of CI (E(LOCI)) of these two approaches which are reported in the Table 2.

For LOG approach, $LOCI = ((\nu^2 - 1)/\nu) \cdot \exp(\bar{W}_{i.}) \cdot \nu = \exp(t_{(n-k), \alpha/2} \sqrt{S_w^2/r})$. Note that, $E(S_w^2) = \frac{1}{k} \sum_{i=1}^k E(S_{wi}^2)$

where $E(S_{wi}^2) = \text{Var}(\log Y)$. The delta method leads to $E(S_w^2) \approx \nu^{-1}$. Using standard results,

$E(\exp(\bar{W}_{i.})) = \lambda_i (\Gamma((1/r) + \nu) / \Gamma \nu)^r$. Although ν and $\bar{W}_{i.}$ need not be independently distributed, the approximate E(LOCI) is obtained and reported in Table 2.

Note that, GLM-LOG and GLM-REC approaches both estimates the expected response by $\bar{Y}_{i.}$ which is an unbiased estimator of μ_i while the estimated response for LOG approach turns out to be the geometric mean (GM) of the relevant group of observations which underestimates the expected responses. Note that, for an unreplicated design, the interval estimates for LOG approach cannot be computed in the absence of an estimate of error variance required therein. Methods like Lenth's method designed especially for analysis of unreplicated factorial experiments could be employed here. However for GLM with known shape parameter, the

confidence limits do not require such an estimate and are computable. The details of the underlying model, estimated responses, confidence limits, CI for the mean response μ_i at

the point X_i (cf. Myers and Montgomery (1997)), LOCI around expected responses and E(LOCI) for the three approaches are summarized in the Table 2.

Table 2: Underlying model and related formulae

Method	Model	Response distribution and estimated response under the respective model	Confidence Limits ^a and LOCI ^b	E(LOCI)
LOG	$W_{ij} = \log(Y_{ij})$ $E\bar{W} = X\bar{\beta}$ $\mu_i = (X'_i\bar{\beta})$	$W_{ij} \sim N(\mu_i, \sigma_w^2)$ $\hat{Y}_{ij} = \exp(\bar{W}_i)$ $= (\prod Y_{ij})^{1/r} = GM$	$\exp(\bar{W}_i \pm t_{(n-k), \alpha/2} \sqrt{S_w^2/r})^a$ $((v^2 - 1)/v) \exp(\bar{W}_i)^b$ $v = \exp(t_{(n-k), \alpha/2} \sqrt{S_w^2/r})$	$\approx E((v^2 - 1)/v) E(\exp(\bar{W}_i))$ $= \frac{\mu_i}{v} \left(\exp(\theta/\sqrt{rv}) - \exp(-\theta/\sqrt{rv}) \right)$ $\left(\Gamma(\frac{1}{r} + \nu)/\Gamma\nu \right)^r, \theta = t_{(n-k), \alpha/2}$
GLM-LOG	$E\bar{Y} = \exp(X'\bar{\beta})$ $\mu_i = \exp(X'_i\bar{\beta})$	$Y_{ij} \sim \text{Gamma}(\lambda_i, \nu)$ $\hat{Y}_{ij} = \bar{Y}_i$	$\bar{Y}_i \exp(\pm Z_{\alpha/2}/\sqrt{rv})^a$ $\bar{Y}_i \left(\exp\{Z_{\alpha/2}/\sqrt{rv}\} - \exp\{-Z_{\alpha/2}/\sqrt{rv}\} \right)^b$	$\mu_i \left(\exp\{Z_{\alpha/2}/\sqrt{rv}\} - \exp\{-Z_{\alpha/2}/\sqrt{rv}\} \right)$
GLM-REC	$E\bar{Y} = (X'\bar{\beta})^{-1}$ $\mu_i = (X'_i\bar{\beta})^{-1}$	$Y_{ij} \sim \text{Gamma}(\lambda_i, \nu)$ $\hat{Y}_{ij} = \bar{Y}_i$	$\bar{Y}_i \left((1 \mp Z_{\alpha/2}/\sqrt{rv})^{-1} \right)^a$ $\bar{Y}_i \left((1 - Z_{\alpha/2}/\sqrt{rv})^{-1} - (1 + Z_{\alpha/2}/\sqrt{rv})^{-1} \right)^b$	$\mu_i \left((1 - Z_{\alpha/2}/\sqrt{rv})^{-1} - (1 + Z_{\alpha/2}/\sqrt{rv})^{-1} \right)$

a: Confidence Limits b: LOCI

2.3 Theoretical Coverages for Three Approaches

Table 3: Underlying model and related formulae

Method	Coverages
GLM-LOG	$G\left(\bar{Y}_i \exp(Z_{\alpha/2}/\sqrt{rv})\right) - G\left(\bar{Y}_i \exp(-Z_{\alpha/2}/\sqrt{rv})\right)$
GLM-REC	$G\left(rv(1 + (1.96/\sqrt{rv})), 1, rv\right) - G\left(rv(1 - (1.96/\sqrt{rv})), 1, rv\right)$
LOG	$L(\underline{\mu}) = 100\left(1 - (1/t_{(n-k), \alpha/2}^2)\right)$

where $G(x, 1, r^v)$ is the cumulative distribution function at x for the $\text{Gamma}(1, r^v)$ distribution. Note that, the coverage probability can be computed for an unreplicated design by taking $r = 1$ for GLM-LOG and GLM-REC approaches.

Unfortunately for fixed α , the L. B. for LOG approach is not sharp due to the approximations employed for its computation. It is decreasing in the number of replications r and number of treatment combinations k through the d. f. associated with the t -quantile. At $\alpha = 0.05$, this equals 87.03 for $r = 2, k = 4$; 81.2 for $r = 3, k = 4$ as well as for $r = 2, k = 8$ and 77.75 for $r = 3, k = 8$.

2.4 Observations on Theoretical Coverages and E(LOCI)

Observations on Theoretical Coverages

The theoretical coverages given for GLM-LOG and LOG

approaches do not attain the desired level of significance. The coverages given for GLM-REC are just above the desired level. The coverages of all the three approaches are well concentrated and increasing in v . The coverages given for GLM-LOG and GLM-REC approaches are increasing in number of replications r but the coverages given for LOG approach are undesirably decreasing in the number of replications r and the number of treatment combinations k .

Observations on Theoretical E(LOCI)

The E(LOCI) for all the three approaches depends only on the value of the particular component μ_i of $\underline{\mu}$. These are increasing in μ_i , decreasing in the number of replications r and the shape parameter v . The theoretical comparison among

the three approaches is not feasible due to the approximations used in the derivation of E(LOCI) and L.B. of LOG approach and is attempted through a simulation study presented in the next section.

3. Simulation study

3.1 Back-ground and Details of the simulation study

Noting the dependence of E(LOCI) for the three approaches on the expected responses $\underline{\mu}$, rather than on the parameters $\underline{\beta}$,

the comparison is based on the values of $\underline{\mu}$. The simulation study is carried out for 2^2 experiments for small number of replications $r=2$ and 3 which are commonly used in industrial experiments. This covers a reasonable representative subset of the parameter space representing a variety of situations that could be encounter in practice used by Patil and Kulkarni (2011). The coefficient matrix is as defined in (1).

For each selected $\underline{\mu}$ the coverages and E(LOCI) for $\nu = 2, 3$ and 5 are simulated and the Box plots based on simulated coverages and simulated E(LOCI) are displayed for comparison for 2 and 3 replications in Figure 1.

3.2 General Observations and Overall Prominent Facts

The simulated coverages for GLM-LOG and GLM-REC approaches exactly conform with the theoretical ones given in Table 3. The simulated coverages for GLM-LOG are below the desired level of significance while for GLM-REC they are just above the desired level. The simulated coverages for LOG approach follow the L.B. given in Table 3 but it is not sharp due to the approximations employed for its computation. The coverages are increasing in ν but are below the desired level. The simulated E(LOCI) of all the three approaches depend only on the value of the particular component μ_i of $\underline{\mu}$. The simulated E(LOCI) for GLM-LOG and GLM-REC approaches exactly match with those given in Table 2. These are increasing in μ_i decreasing in the number of replications r , and the shape parameter ν . The simulated E(LOCI) for LOG approach are larger than the approximation reported in Table 2. These are increasing in the particular component μ_i , decreasing in the number of replications r and the shape parameter ν .

3.3 The Summary of Overall Behavior of the Three Approaches

- The simulated coverages for all the three approaches are well-concentrated but below the desired level for LOG and GLM-LOG approaches and simulated coverages are above the desired level for GLM-REC approach.
- The coverages for GLM-LOG approach are uniformly inferior to those of LOG approach for two replications and superior for more replications.
- The E(LOCI) for GLM-REC are largest among the three approaches. For GLM-LOG approach they are uniformly smaller than LOG approach.

Although Lewis, Montgomery and Myers (2001b) advocated the use of GLM in the analysis of factorial designs, here it is

noted that, the LOG approach outperforms GLM-LOG for two replications with respect to coverage probabilities which is typically encountered situation in industrial experiments. However the extent of outperformance does not attain the desired level.

In the next section, we suggest a modification to CI for GLM-LOG and LOG approach that shows decisive improvement to attain desired confidence coefficient.

4. New proposals

A modification to GLM-LOG based CI which is a kind of 're-centering' of the existing CI is suggested to capture the region of high density and is referred to as the 'M-GLM' approach in the sequel.

4.1 M-GLM approach

The M-GLM approach is designed to achieve the coverage probability exactly equal to the desired confidence coefficient by a sort of 're-centering' the CI for GLM-LOG approach. For the given level α , a number ' δ ' can be selected so that,

$$Pr\left(r\nu \exp(-\delta Z_{\alpha/2}/\sqrt{r\nu}) \leq \sum_{j=1}^r Y_{ij}/\mu_i \leq r\nu \exp(\delta Z_{\alpha/2}/\sqrt{r\nu})\right) = 1 - \alpha \quad (4)$$

The resulting CI limits are $\bar{Y}_i \cdot (\exp(\pm \delta Z_{\alpha/2}/\sqrt{r\nu}))$ and the corresponding E(LOCI)

$$E(LOCI) = \mu_i \cdot (\exp(\delta Z_{\alpha/2}/\sqrt{r\nu}) - \exp(-\delta Z_{\alpha/2}/\sqrt{r\nu})) \quad (5)$$

Note that the modification is effective for small shape parameters and is not required for large shape parameters.

4.2 BC-LOG approach

For LOG approach, $W = \log(Y)$. Using the delta method

leads to $\mu_w = E(\log Y_i) \approx \log(\mu_i) - \frac{1}{2\nu}$ and

$g^{-1}(\mu_w) = \exp(\mu_w) \approx \mu_i \exp\left(-\frac{1}{2\nu}\right)$. Thus, $L_y = \exp(L_w)$

and $U_y = \exp(U_w)$ are the limits for $\mu_i \exp\left(-\frac{1}{2\nu}\right)$ and not

for μ_i . This introduces a large bias particularly for small shape parameter. The suggested bias corrected CI

$$\left(L_y \exp\left(\frac{1}{2\nu}\right), U_y \exp\left(\frac{1}{2\nu}\right)\right) \quad (6)$$

and the associated LOCI is $[\exp(\frac{1}{2\nu})(U_y - L_y)]$.

The Coverages and E(LOCI) for M-GLM and BC-LOG approach are simulated by using the method discussed in Section 3 for 2^2 experiments for $\nu = 2, 3$ and 5. The box plots together with those for its uncorrected version for the sake of direct comparison are displayed in Figure 2 for $r = 2$ and 3 replications.

Although the coverage probabilities of these CI are improved,

for small shape parameter, these lengths are larger than the uncorrected ones. Similar simulation study for 2^3 full saturated

factorial experiments with $r = 2$ and 3 replications is also carried out. This study also revealed similar conclusions.

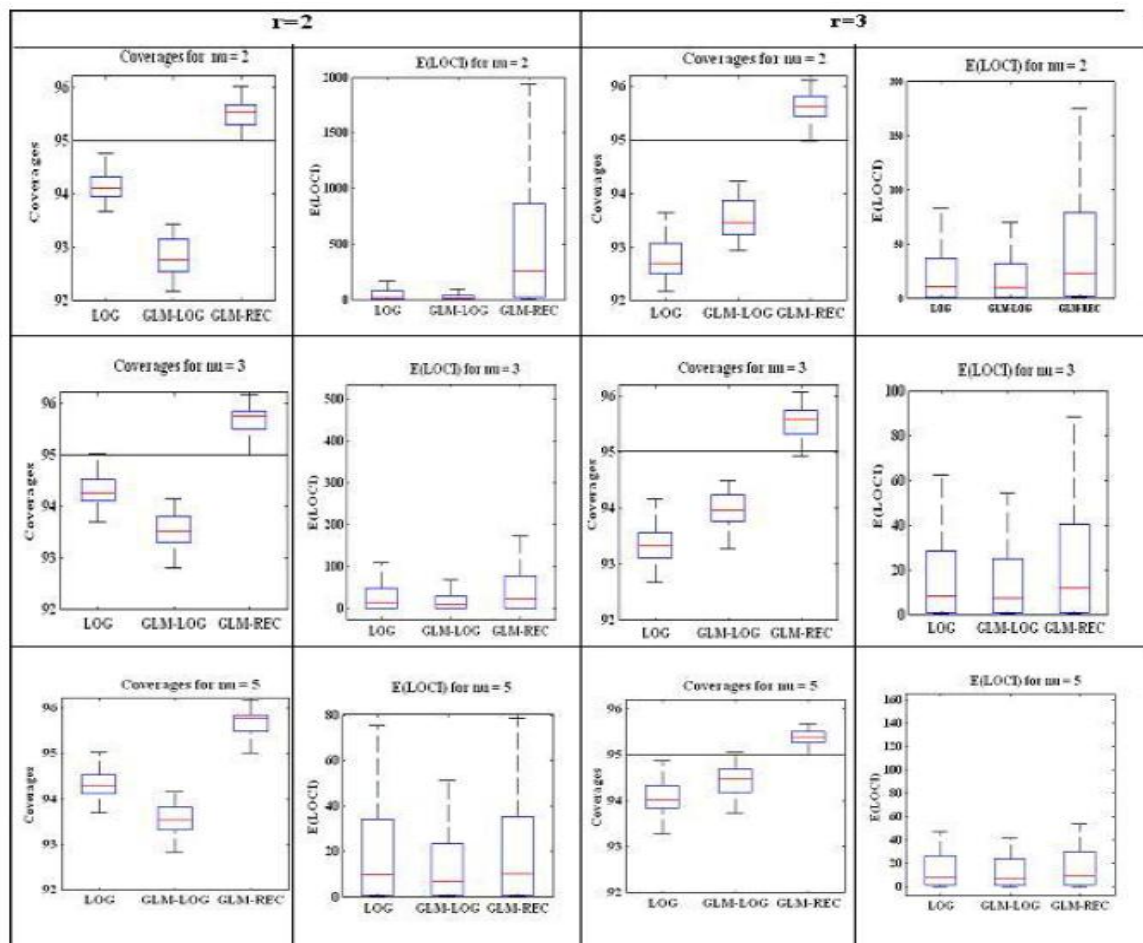
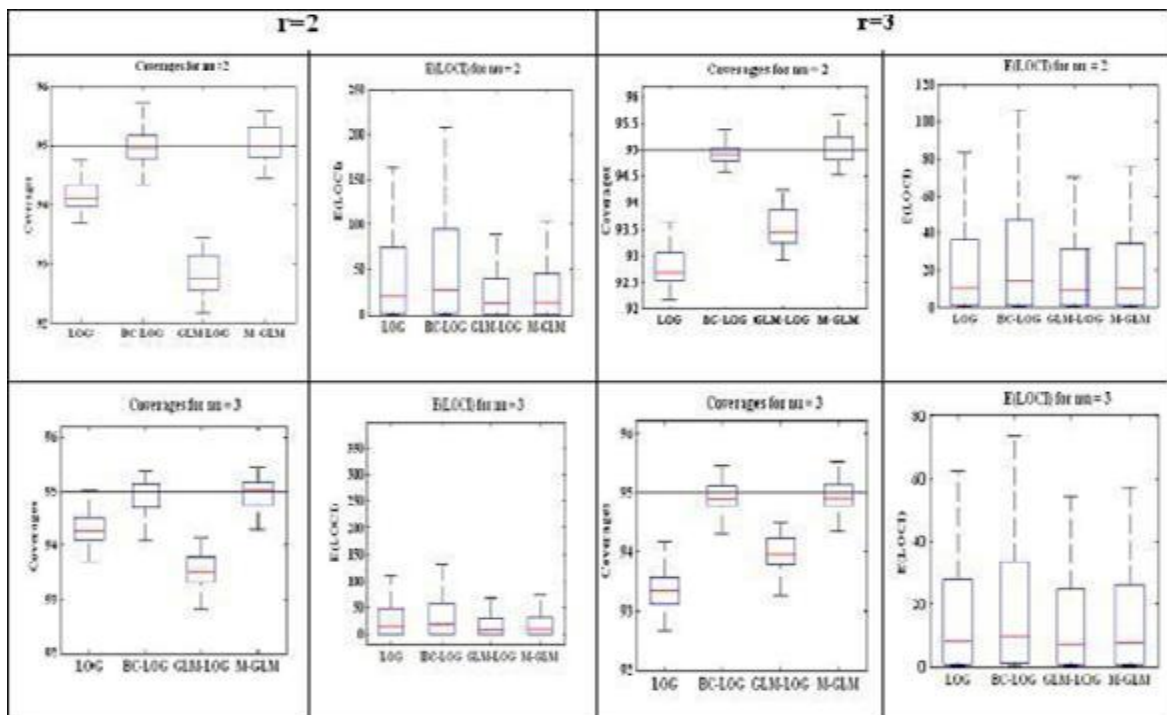


Fig 1: Box plots of simulated coverage's and E(LOCI) for three approaches for 2^2 factorial experiments



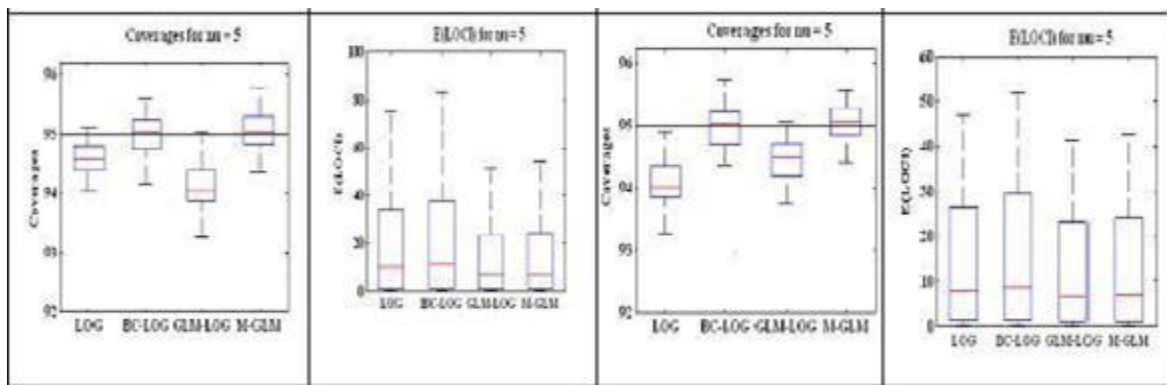


Fig 2: Box plots of simulated coverages and E(LOCI) for GLM-LOG and LOG approaches along with their modified versions for 2^2 factorial experiments

5. Concluding remarks

For interval estimation of expected response with small replications, GLM-REC approach has largest coverage probability amongst the three approaches but with largest expected length. LOG approach has larger coverage probability for small shape parameter ($\nu = 2$) than GLM-LOG but still little below the desired level and with larger expected lengths. The suggested BC-LOG improves the coverage probabilities of LOG approach in this case but at the cost of increased expected lengths. The suggested M-GLM CI resolves this problem at the cost of a little increased expected length than GLM-LOG approach and is strongly recommended.

6. Acknowledgement

This work is funded by BCUD, Savitribai Phule Pune University, Pune (14SCI001250).

7. References

1. Lewis S, Montgomery D, Myers R. 'Examples of Designed Experiments with Non-normal Responses'. *Journal of Quality Technology*. 2001a; 33:265-278.
2. Lewis S, Montgomery D, Myers R. 'CI Coverage for Designed Experiments Analyzed with GLMs'. *Journal of Quality Technology*. 2001b; 33:279-291.
3. McCullagh P, Nelder J. 'Generalized Linear Models', *Chapman and Hall, London*, 1989.
4. Myers R, Montgomery D. A Tutorial on Generalized Linear Models'. *Journal of Quality Technology*. 1997; 29:274-290.
5. Nelder J, Wedderburn R. 'Generalized Linear Models'. *J. R. Statistics Soc. A*. 1972; 135:370-384.
6. Patil S, Kulkarni H. Analysis of 2^n Factorial Experiments with Exponentially Distributed Response Variable, *Applied Mathematical Sciences*, 2011; 5(10):459-476.