Special Proceedings of the 22nd Annual Conference of SSCA held at S.P. Pune University,
Pune, during 02-04 January 2020, 145-152

Selection of Designs for Model Misspecification in Generalized Linear Models: A Review

Siuli Mukhopadhyay[1] **and Ishapathik Das**[2]
[1]*Department of Mathematics, Indian Institute of Technology Bombay, India*
[2]*Department of Mathematics and Statistics, Indian Institute of Technology Tirupati, India*

Received: 10 July 2020; Revised: 20 August 2020; Accepted: 21 August 2020

Abstract

This article provides a brief review on design selection under model misspecification in linear and generalized linear models. Design selection for fitting a hypothesized model is one of the main focuses of response surface methodology. However, if the assumptions regarding the relationship between the response and the covariates are incorrect, then the design based on the assumed model may not provide accurate results. In generalized linear models (GLMs), a certain form of the linear predictor and the link function of the model is usually assumed and the selected designs are based on these assumptions. Model misspecification in GLMs can arise when the form of the linear predictor and/or the link function assumed is not correct. Many researchers have proposed several methods for selecting appropriate designs accounting for model bias and the prediction variance for both linear and generalized linear models. The literature review presented here discusses several existing methods in the literature based on the mean squared error criterion for comparing/selecting designs robust to the possible misspecification in the model. Several papers based on robust designs for GLMs are highlighted here. The method of comparing designs by quantile dispersion graphs (QDGs) approach addressing the linear predictor misspecification problem using an unknown function and the link function misspecification problem using a family of link functions is discussed in detail. A numerical example based on real data is provided to illustrate the QDGs methodology.

Key words: Family of link functions; Mean squared error of prediction; Quantile dispersion graphs; Robust designs.

1. Introduction

One of the main purposes of response surface methodology (RSM) is to choose an appropriate design for fitting a hypothesized model. Usually a low-degree polynomial or a simple linear model is used to explain the complex and possibly non linear relationship between the response variable and the inputs/covariates. Since the simple fitted model may not adequately approximate the unknown functional relationship that depicts the true mean

Corresponding Author: Siuli Mukhopadhyay
Email: siuli@math.iitb.ac.in

response, there is always a chance of estimates being biased. Thus, giving rise to the model misspecification problem. Due to this reason the chosen design should protect against the possibility of a sizeable model bias. Box and Draper (Box & Draper (1959) and Box & Draper (1963)) introduced the so-called integrated mean squared error (IMSE) criterion which accounts for both prediction variance and model bias and advised experimenters to choose designs on the basis of the IMSE. Instead of looking at an overall measure like the average MSE, Giovannitti-Jensen & Myers (1989) and Vining & Myers (1991) used a graphical approach to evaluate how a design performs over every portion of the region of interest in terms of IMSE. More recently, Mukhopadhyay & Khuri (2008) presented the technique of qunatile plots for evaluating and comparing response surface designs on the basis of the mean squared error of prediction (MSEP). Four MSEP-related criteria functions free of any unknown parameters that pertain to the unfitted true model and error variance were proposed. They obtained plots of the quantiles of these criterion functions on concentric spheres within a region of interest. These quantile plots gave complete information concerning the distribution of each criterion function over the selected spheres.

Recently, there has been an increase in interest among researchers to study designs robust to model misspecification in the context of generalized linear models (GLMs). Model misspecification in GLMs is a little more complex than in linear models, since in GLMs along with simple form of the linear predictor, the experimenter also assumes a form for the link function. If the assumptions regarding the functional form of the linear predictor or/and the link function are incorrect, then the inference drawn from the fitted model may not provide accurate results, giving rise to model misspecification problem in GLMs.

Selecting robust designs for GLMs have been studied by Abdelbasit & Butler (2006), Woods et al. (2006) and Dror & Steinberg (2006). In the context of logistic regression models, Adewale & Wiens (2009) used the average mean-squared error criterion to generate designs less sensitive to possible misspecifications in the linear predictors. Their work was extended by Adewale & Xu (2010) where misspecification in both linear predictors and link functions were considered. More recently, Mukhopadhyay & Khuri (2012) used quantile dispersion graphs based on MSEP to compare designs for GLMs in the presence of model misspecification in linear predictors. Their approach accounted for the bias of the fitted model's parameter estimates in addition to their variances.

2. Model Misspecification in GLMs

GLMs are usually specified by three components:

- Distributional component: It is assumed that the data of size n $y_1, \ldots, y_n$, are independent and have the following density function,

$$s(y_j|\theta_j, \phi) = \exp\left[\frac{y_j\theta_j - b(\theta_j)}{a(\phi)} + c(y_j, \phi)\right], \; j = 1, \ldots, n, \tag{1}$$

where $b(\cdot)$, $c(\cdot)$ are known functions and ϕ is the unknown dispersion parameter. The mean and variance of y_j are, $E(y_j) = \mu_j = \frac{db(\theta_j)}{d\theta_j}$ and Var $(y_j) = \sigma_j^2 = a(\phi)\frac{d^2b(\theta_j)}{d\theta_j^2}$, respectively.

- Linear Predictor: The linear predictor, denoted by $\eta(\mathbf{x})$, is a function of the p control variables $\mathbf{x} = (x_1, \dots, x_p)^T$.

- Link function: The linear predictor $\eta(\mathbf{x})$ is related to the mean response, $\mu(\mathbf{x})$, through a link function g, where the inverse of g, denoted by h, is assumed to exist. The true relationship between η and $\mathbf{x}$ being usually unknown and of highly nonlinear nature.

As mentioned above, the true relationship between $\eta(\mathbf{x})$ and the vector $\mathbf{x}$ of control variables is usually unknown. The experimenter approximates the unknown relationship by a low-order polynomial model of the form,

$$\eta(\mathbf{x}) = \mathbf{z}^T(\mathbf{x})\boldsymbol{\beta}, \tag{2}$$

where, $\mathbf{z}^T(\mathbf{x})$ is a known vector function of $\mathbf{x}$ and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. Under the assumed model, the estimated mean response is,

$$\hat{\boldsymbol{\mu}}(\mathbf{x}) = h[\hat{\eta}(\mathbf{x})] = h[\mathbf{z}^T(\mathbf{x})\hat{\boldsymbol{\beta}}], \tag{3}$$

where $\hat{\boldsymbol{\beta}}$ is a maximum likelihood estimate of $\boldsymbol{\beta}$. However, suppose the true functional form of the linear predictor is different from the fitted form and is actually,

$$\eta_T(\mathbf{x}) = \mathbf{z}^T(\mathbf{x})\boldsymbol{\beta} + f(\mathbf{x}), \tag{4}$$

where $f(\mathbf{x})$ is not known and the true mean response is,

$$\mu_T(\mathbf{x}) = h[\eta_T(\mathbf{x})] = h[\mathbf{z}^T(\mathbf{x})\boldsymbol{\beta} + f(\mathbf{x})]. \tag{5}$$

The MSEP for the estimated mean response when the linear predictor is misspecified from Mukhopadhyay & Khuri (2012) is given by

$$\begin{aligned} \text{MSEP}[\hat{\mu}(\mathbf{x})] &\doteq \left[\frac{dh[\eta(\mathbf{x})]}{d\eta(\mathbf{x})} + f(\mathbf{x})\frac{d^2h[\eta(\mathbf{x})]}{d\eta^2(\mathbf{x})}\right]^2 \text{ Var } [\hat{\eta}(\mathbf{x})] \\ &+ \left\{\text{Bias } [\hat{\eta}(\mathbf{x})]\left[\frac{dh[\eta(\mathbf{x})]}{d\eta(\mathbf{x})} + f(\mathbf{x})\frac{d^2h[\eta(\mathbf{x})]}{d\eta^2(\mathbf{x})}\right]\right\}^2, \end{aligned} \tag{6}$$

where,

$$\begin{aligned} \text{Bias } [\hat{\eta}(\mathbf{x})] &= E[\hat{\eta}(\mathbf{x})] - \eta_T(\mathbf{x}) = \mathbf{z}^T(\mathbf{x})E(\hat{\boldsymbol{\beta}}) - \mathbf{z}^T(\mathbf{x})\boldsymbol{\beta} - f(\mathbf{x}) \\ &= \mathbf{z}^T(\mathbf{x}) \text{ Bias } (\hat{\boldsymbol{\beta}}) - f(\mathbf{x}), \end{aligned}$$

and

$$\text{Var } [\hat{\eta}(\mathbf{x})] = \mathbf{z}^T(\mathbf{x}) \text{ Var } (\hat{\boldsymbol{\beta}})\mathbf{z}(\mathbf{x}).$$

The bias and variance of $\hat{\boldsymbol{\beta}}$ under a misspecified linear predictor are

$$\text{Bias }(\hat{\boldsymbol{\beta}}) \doteq \mathbf{H}_n^{-1}\mathbf{b}, \tag{7}$$

and

$$\text{Var }(\hat{\boldsymbol{\beta}}) \doteq \frac{1}{N}\mathbf{H}_n^{-1}\tilde{\mathbf{H}}_n\mathbf{H}_n^{-1}, \tag{8}$$

where $\mathbf{H}_n = \mathbf{Z}^T\mathbf{PWZ}$, $\tilde{\mathbf{H}}_n = \mathbf{Z}^T\mathbf{PW}_T\mathbf{Z}$, and $\mathbf{b} = \frac{\mathbf{Z}^T\mathbf{P}(\boldsymbol{\mu}_T-\boldsymbol{\mu})}{a(\phi)}$; $\mathbf{Z}$ is a matrix with rows $\mathbf{z}^T(\mathbf{x}_j)$, $j = 1,\ldots,n$ and $\mathbf{P}$ is an $n \times n$ diagonal matrix with elements $\frac{n_j}{N}$ with N is the total number of observations, i.e., $N = \sum_{j=1}^{n} n_j$. Both, $\mathbf{W}$ and $\mathbf{W}_T$ are $n \times n$ diagonal matrices with elements, $w_j\,(=\frac{d\mu_j/d\eta_j}{a(\phi)})$, $w_{T,j}\,(=\frac{\text{Var }(y_j)}{a^2(\phi)})$, respectively, where Var (y_j) is the true variance of y_j.

A scaled version of the MSEP (SMSEP) was used for design comparison. Their main goal was to select designs with lower values of SMSEP. For comparing two designs say D_1 and D_2, if the SMSEP of D_1 was lower than D_2 then design D_1 was said to have better prediction capability than D_2. Thus, implying that the predictive performance of design D_1 is more robust to misspecification in the linear predictor than D_2. However, two major difficulties in using the SMSEP as a design criterion was its dependency on the unknown model parameters and $f(\mathbf{x})$. Mukhopadhyay & Khuri (2012) addressed the linear predictor misspecification problem by an unknown function which was estimated using parametric empirical kriging at any point in the design region. The dependence of SMSEP on the model parameters was answered by the quantile dispersion graphs (QDGs) approach.

Das et al. (2015) considered robust GLM designs for misspecification in both linear predictors and link functions. To address the possibility of incorrect forms of link functions, they used the works of Prentice (1976); Pregibon (1980); Aranda-Ordaz (1981); Guerrero & Johnson (1982); Stukel (1988); Czado (1989, 1997) on generalized family of link functions for GLMs.

A family of parametric link functions were defined, relating $\eta(\mathbf{x})$ and $\mu(\mathbf{x})$ by $\mu = E(y|\mathbf{x}) = h(\boldsymbol{\alpha},\eta)$, where $h(\boldsymbol{\alpha},\cdot)$ is inverse of the parametric link function parameterized by $\boldsymbol{\alpha}$ the link parameter vector (Czado, 1997). Thus, for misspecification in both the linear predictor as well as the link function, the assumed model is

$$\mu(\mathbf{x}) = h[\boldsymbol{\alpha}_0, \eta(\mathbf{x})],$$

where $h(\boldsymbol{\alpha}_0,\cdot)$ is the assumed link function belonging to the family $\boldsymbol{\Lambda} = \{h(\boldsymbol{\alpha},\cdot) : \boldsymbol{\alpha} \in \boldsymbol{\Omega}\}$, and the true model

$$\mu_T(\mathbf{x}) = h[\boldsymbol{\alpha}_T, \eta_T(\mathbf{x})],$$

is

$$\eta_T(\mathbf{x}) = \mathbf{Z}(\mathbf{x})\boldsymbol{\beta} + f(\mathbf{x})$$

is the true linear predictor and $\boldsymbol{\alpha}_T$ the true link parameter. The MSEP of $\hat{\boldsymbol{\mu}}(\mathbf{x})$ from Das et al. (2015) is given by

$$MSEP[\hat{\boldsymbol{\mu}}(\mathbf{x})] = Var[\hat{\boldsymbol{\mu}}(\mathbf{x})] + Bias[\hat{\boldsymbol{\mu}}(\mathbf{x})][Bias\{\hat{\boldsymbol{\mu}}(\mathbf{x})\}]^T,$$

where

$$Var[\hat{\boldsymbol{\mu}}(\mathbf{x})] \;=\; \left[\frac{\partial h}{\partial \eta}\right]_{(\boldsymbol{\alpha}_0,\eta_T(\mathbf{x}))} \mathbf{Z}(\mathbf{x}) Var(\hat{\boldsymbol{\beta}})\mathbf{Z}^T(\mathbf{x}) \left[\frac{\partial h}{\partial \eta}\right]^T_{(\boldsymbol{\alpha}_0,\eta_T(\mathbf{x}))},$$

and

$$\begin{aligned} Bias[\hat{\boldsymbol{\mu}}(\mathbf{x})] \;&=\; \left[\frac{\partial h}{\partial \boldsymbol{\alpha}}\right]_{(\boldsymbol{\alpha}_T,\eta_T(\mathbf{x})} (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_T) \\ &+\; \left[\frac{\partial h}{\partial \eta}\right]_{(\boldsymbol{\alpha}_0,\eta_T(\mathbf{x}))} [\mathbf{Z}(\mathbf{x}) Bias(\hat{\boldsymbol{\beta}}) - f(\mathbf{x})], \end{aligned}$$

The asymptotic bias and variance of $\hat{\boldsymbol{\beta}}$ are given by

$$\begin{aligned} Bias(\hat{\boldsymbol{\beta}}) \;&=\; \mathbf{H}_n^{-1} b, \text{ and} \\ Var(\hat{\boldsymbol{\beta}}) \;&=\; \frac{1}{N}\mathbf{H}_n^{-1}\tilde{\mathbf{H}}_n\mathbf{H}_n^{-1}, \end{aligned}$$

where

$$b = \sum_{i=1}^{n} \frac{1}{N}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}[Var(y_i)]^{-1}(\mu_{T,i} - \mu_i),$$

$$\tilde{\mathbf{H}}_n \;=\; \frac{1}{N}\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}[Var(y_i)]^{-1}[Var(y_{T,i})][Var(y_i)]^{-1}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T},$$

and

$$\mathbf{H}_n \;=\; \frac{1}{N}\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}[Var(y_i)]^{-1}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T} - \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{q}\frac{\partial^2 \theta_{ij}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}(y_{ij} - \mu_{ij})n_i.$$

See Das et al. (2015) for details.

2.1. Example

We consider a real data set (Calandra Granaria data (Adewale & Xu, 2010)) containing information about studying the mortality of grain beetle after exposure to ethylene oxide (C_2H_4O). This same example was considered in Das et al. (2015). The response variable y is the proportion of killed grain beetle after one-hour exposure of 10 different levels of concentrations of C_2H_4O, which is considered as the explanatory variable (x) of the model. Here, we compare three designs: (i) the design D_7 (original design), (ii) the design D_8 ("Naive" design) and (iii) the regular optimal design D_9 (Adewale & Xu, 2010) under a misspecified linear predictor. The data set and design settings can be found in Table 7 of Das et al. (2015).

We start fitting the data with the linear predictor

$$\eta(x) = \beta_0 + \beta_1 x, \tag{9}$$

and use the logistic link function. After estimating the unknown parameters of the model using maximum likelihood estimation method, we see $\hat{\beta}_0 = -3.4429$ and $\hat{\beta}_1 = 14.4404$, and the deviance of the fitted model is 36.2498 with 8 degrees of freedom. The observed information provides that the P-value is less than 0.0001, showing a lack of fit due to the possible misspecification of the linear predictor of the model. So, we add an unknown function to the previous linear predictor to have the modified linear predictor as

$$\eta(x) = \beta_0 + \beta_1 x + f(x). \tag{10}$$

The values of the unknown function f are first estimated at design points using the method given in Section 4.1 of Das et al. (2015) and then estimated at any points other than the design points using parametric empirical kriging. We see that the deviance of the fitted model after adding an unknown function f is decreased to 4.9778, showing an improvement of the fit by addressing the linear predictor misspecification of the model.

For comparing the performance of three designs concerning the proximity to the center/boundary of the design region, the experimental region $\mathcal{R} = \{x : 0.0330 \leq x \leq 0.3940\}$ is divided into several concentric regions $\mathcal{R}_\nu$ parametrized by some parameter $\nu \in [0.5, 1]$. The designs are compared based on the minimum and maximum quantiles of the estimated MSEP values over randomly selected 1000 samples from $\mathcal{R}_\nu$ and 1000 samples from a 95% confidence region $\mathcal{C}$ of the regression parameter vector $\boldsymbol{\beta}$. The minimum and maximum quantiles of the three designs for $\nu = 0.6, 0.7, 0.8, 0.9$, are shown in Figure 1, which is known as the quantile dispersion graphs (QDGs). From the QDGs, we see that the minimum quantiles are close to each other for all designs. The maximum quantiles of D_9 are larger than D_7 and D_8 if $\mathfrak{p} > 0.5$ for all values of ν. So, the prediction capabilities of D_7 and D_8 are better than the design D_9, while designs D_7 and D_8 have comparable prediction capabilities throughout the region as the maximum quantiles are very close to each other for all values of ν. It can also be noted by observing the differences of maximum and minimum quantiles of the designs that D_7 and D_8 are more robust than D_9 with respect to the changes of the values of $\boldsymbol{\beta}$. More details about this example can be found in Section 5.3 of Das et al. (2015).

3. Some New Directions

Though the topic of model misspecification and its effect on design selection has been discussed by several researchers for single response linear and generalized linear models, very little work has been done in the multivariate response case. However, in many experimental situations, instead of one response, several such responses are recorded for the same subject. This is very common in drug testing experiments where along with efficacy of the drug, toxic effect of the drug are also measured, and the two responses are then modeled using a bivariate distribution. Very recently, Das & Mukhopadhyay (2019) discussed the effect

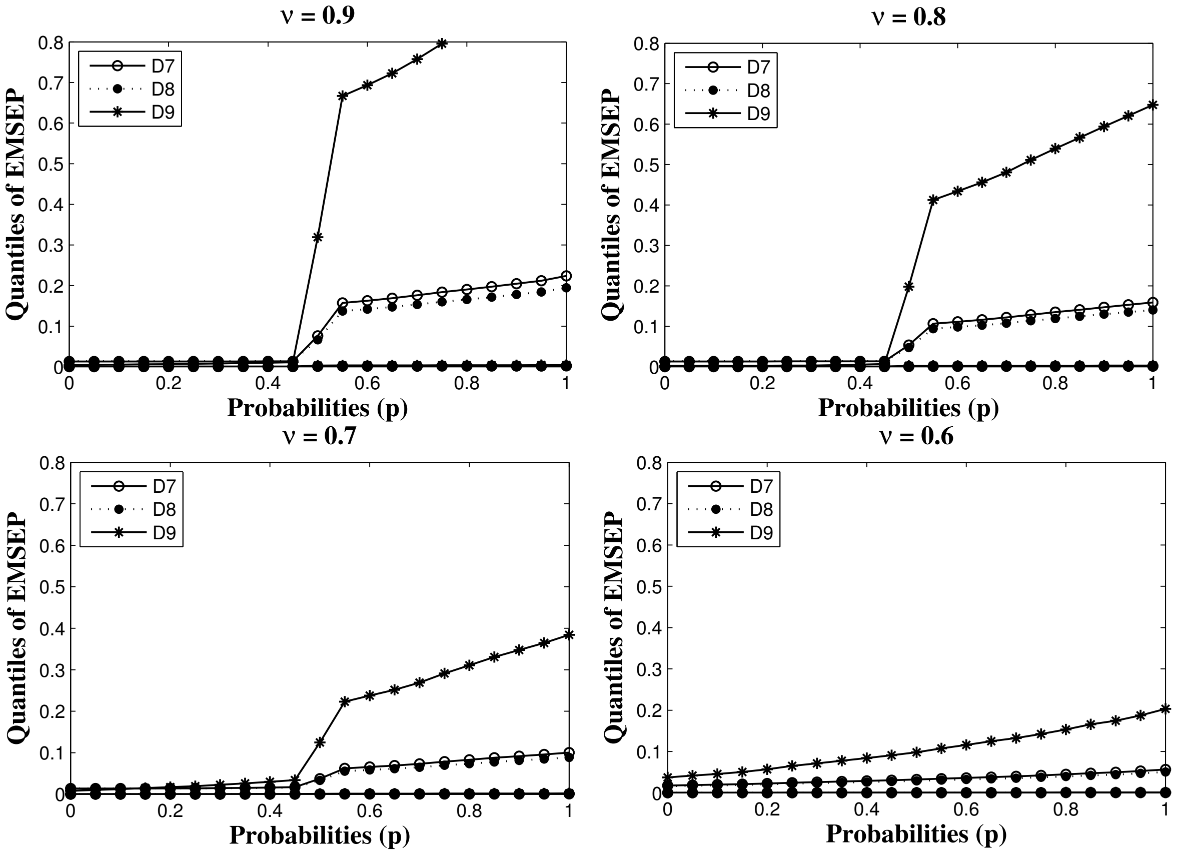

Figure 1: Quantile dispersion graphs for designs D_7, D_8, and D_9. This figure is reproduced from Figure 3 of Das et al. (2015).

of such misspecification on design selection for multinomial GLMs and proposed the use of quantile dispersion graphs to select robust designs. While multivariate kriging was used to tackle the unknown functional relationships between the linear predictors and covariates, a parametric link function family for the multinomial distribution (Das & Mukhopadhyay (2014)) was used for possible link function correction. Compromised exact D- optimal designs which are robust to possible misspecifications in the model and link functions were discussed recently by Singh & Mukhopadhyay (2019) for gene sequence studies modeled by count time series models.

References

Abdelbasit, K. M. and Butler, N. A. (2006). Minimum bias designs for generalized linear models. *Sankhya: The Indian Journal of Statistics (2003-2007)*, **68(4)**, 587–599.

Adewale, A. J. and Wiens, D. P. (2009). Robust designs for misspecified logistic models. *Journal of Statistical Planning and Inference*, **139**, 3–15.

Adewale, A. J. and Xu, X. (2010). Robust designs for generalized linear models with possible overdispersion and misspecified link functions. *Computational Statistics and Data Analysis*, **54**, 875–890.

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, **61**, 357–363.

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. *Journal of the American Statistical Association*, **54**, 622–654.

Box, G. E. P. and Draper, N. R. (1963). The choice of a second order rotatable design. *Biometrika*, **50**, 335–352.

Czado, C. (1989). *Link misspecification and data selected transformations in binary regression models*. Technical report, Ph.D. Thesis. School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY.

Czado, C. (1997). On selecting parametric link transformation families in generalized linear models. *Journal of Statistical Planning and inference*, **61**, 125–139.

Das, I., Aggarwal, M. and Mukhopadhyay, S. (2015). Robust designs in generalized linear models: A quantile dispersion graphs approach. *Communications in Statistics-Simulation and Computation*, **44(9)**, 2348–2370.

Das, I. and Mukhopadhyay, S. (2014). On generalized multinomial models and joint percentile estimation. *Journal of Statistical Planning and Inference*, **145**, 190–203.

Das, I. and Mukhopadhyay, S. (2019). Robust designs for multinomial models. *Communications in Statistics-Simulation and Computation*, **48(10)**, 2998–3021.

Dror, H. A. and Steinberg, D. M. (2006). Robust experimental design for multivariate generalized linear models. *Technometrics*, **48**, 520–529.

Giovannitti-Jensen, A. and Myers, R. H. (1989). Graphical assessment of the prediction capability of response surface designs. *Technometrics*, **31**, 159–171.

Guerrero, V. and Johnson, R. (1982). Use of the box cox transformation with binary response models. *Biometrika*, **69**, 309–314.

Mukhopadhyay, S. and Khuri, A. I. (2008). A new graphical approach for comparing response surface designs on the basis of the mean squared error of prediction criterion. *Statistics and Applications*, **6**, 293–324.

Mukhopadhyay, S. and Khuri, A. I. (2012). Comparison of designs for generalized linear models under model misspecification. *Statistical Methodology*, **9**, 285–304.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society*, **29**, 15–24.

Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics*, **32**, 761–768.

Singh, R. and Mukhopadhyay, S. (2019). Exact bayesian designs for count time series. *Computational Statistics and Data Analysis*, **134**, 157 – 170.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, **83**, 426–431.

Vining, G. G. and Myers, R. H. (1991). A graphical approach for evaluating response surface designs in terms of the mean squared error of prediction. *Technometrics*, **33**, 315–326.

Woods, D. C., Lewis, S. M., Eccleston, J. A. and Russell, K. G. (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics*, **48**, 284–292.