



## Performance Comparison between CERES and PR Plots in Detecting the Heteroscedasticity Problem using Liver Cancer and Simulated Data

<sup>1</sup>Nasir Saleem, <sup>2</sup>Dr. Atif Akbar, <sup>3</sup>Prof. Madya Dr. Shamshuritawati binti Sharif, <sup>4</sup>Prof. Dr. Abu Sayed Md Al Mamun, <sup>5</sup> Prof. Dr. A. H. M. Rahmatullah Imon, <sup>6</sup>Dr. Shakeel Ahmad

<sup>1</sup>Mphil (Statistics) Department of Statistics Bahauddin Zakariya university(BZU), Multan, Pakistan

<sup>2</sup>Associated professor Department of Statistics Bahauddin Zakariya university (BZU), Multan, Pakistan

<sup>3</sup>School of Quantitative sciences Universiti Utara Malaysia

<sup>4</sup>Department of Statistics Rajshahi University, Bangladesh

<sup>5</sup>Department of mathematical sciences, Ball State university, (USA)

<sup>6</sup>Lecturer Department of Statistics Bahauddin Zakariya university (BZU), Multan, Pakistan



### Article History

Received: 14/04/2022

Accepted: 23/04/2022

Published: 27/04/2022

Vol – 1 Issue – 2

PP: - 43-57

### Abstract:

Binomial regression model is a very important type of generalized linear model (GLM) and has been applied in wide area such as Liver cancer data. In diagnosing the model, we need to check for the heteroscedasticity problem because it happens when the standard errors of a variable are non-constant. heteroscedasticity will impact the validity and abuse the assumptions of binomial regression. Commonly, conditional expectations and residuals (CERES) and partial residual (PR) plots have been implemented for the identification of heteroscedasticity in the data set. In this paper, we present the comparison analysis between CERES and PR plots by using two different kind of data which are liver cancer data and simulated data. In simulation, we have selected four different sample sizes which are 25, 50, 100, and 200, and variance of error term,  $\gamma = 0.00, 0.15, 0.38, 0.75$ . After that, a 10,000 simulated data is conducted using the R software. The results show the PR plots can detect heteroscedasticity better than CERES plots as due to the larger disperity and the PR plots gives better visual diagnostic for heteroscedasticity as compare to CERES plots. As a summary, this research found that PR plots is the appropriate solution for checking the heteroscedasticity.

**Keywords:** CERES plots, Partial residual plots, Heteroscedasticity

## 1. Introduction

Statistical methods have wide range of application in liver cancer research investigation (Lukman et al. 2019) and (Liu et al. 2019). The identification of relationships among a set of elements is the major concern of binomial regression analysis. Regression diagnostics is the basic requirement to apply regression analysis to reach reliable conclusions. It is necessary to apply regression diagnostics to avail of reliable conclusions. In scientific investigations, it is appealing to develop such methods that have wide applicability with computational ease.

The statistical graphics are most important for the streams analyze data. In our daily life visual representation of data sets through graphs, play an important role in all activities. Before applying regression analysis in liver cancer data and other applied sciences, the violation of assumptions should be addressed Homoscedasticity and heteroscedasticity are some of the important assumptions in regression modeling (Montgomery et al. 2021). The generalized linear model (GLM) (Gill, 2000) is an extended forms of the linear regression model with the response variable follows some

exponential family of distributions expects normal. The three popular components for the GLM are random, systematic, and link function. Nelder and Wedderburn (1972) proposed the idea of class of GLM and establish many models. We start study models are GLM with the binomial regression. The iterative weight leas square method (IWLS) is an estimation method is used to estimate of parameter. In this paper, we would like to consider a binomial response which has wide applications in modeling Liver cancer and many other types of data. The estimation of GLM parameters and their optimality heavily depend on some standard assumptions. Diagnostics are designed to find problems with the assumptions of any statistical procedure.

### 1.2 Review on CERES and Partial residual plots

Partial residual (PR) plots, first introduced by Ezekiel (1924), represent a graphical construction for observing the direction and extent of the linearity of a regressor variable. It has a long history for its prominence. Larsen and McCleary (1972) should get credit for the name PR plots. Wood (1973) referred to them as compared plus residual plots. Mallows (1986) extended these first-order plots to higher orders, the so-called augmented partial residual plots. Mansfield and Conerly (1987) consider the expectation properties of based

PR plots by obtaining algebraic representations of using the true model distributional properties. Cook (1993) obtained further theoretical underpinnings of these plots and proposed and extended class, the CERES plots. CERES is an abbreviated acronym for 'combing conditional expectations and residuals. PR plots also called component plus residual plots. The properties of PR plots were systematically explored by Cook (1993) and Cook and Croos-Dabrera (1998). Berk and Booth (1995) compare PR plots with several other diagnostic plots. Fowlkes (1987) suggested an adaption of PR plots for logistic regression. Landwehr (1983) suggested and application of PR plots for logistic regression. Our general goal, which we make more specific a bit later, is to investigate conditions under which CERES and PR plots can provide useful information in class of GLM. Fowlkes (1987) and Landwehret al. (1984) argued that PRplots may be useful for assessing nonlinearity in binary logistic regression. Landwehr and Pregibon (1993) study these plots for GLMs under canonical links. Kahng and Lee (2004) discussed the usefulness of CERES plots in GLMs. Park and Hastie (2007) discussed the technique of algorithm for regularized the GLMs. Imran and Akbar (2020) discussed the construction of partial residuals using response residuals for the inverse Gaussian regression model is carried out to explore structure and usefulness for visualizing diagnostics the outliers, multicollinearity, and heteroscedasticity and curvature as a function of selected predictors. The practical implementation of these plots can be seen in many fields Wouters et al. (2018).

In this study, CERES and PR plots are constructed for binomial regression model (BRM). These plots provide suitable diagnostics for model specification. Applied scientists usually require simple, powerful, and wide applicable techniques with computational ease. It is cumbersome to practitioners to learn and apply computationally intensive statistical methods. This study explore such an idea while offering the importance of CERES and PR plots in regression diagnostics without applying the conventional tests. To assess the diagnostic value of CERES and partial residual plots in binomial regression addressing violation of assumptions via real and simulated data. Use CERES and partial residual plots for the detection of heteroscedasticity. We also made the comparison of CERES and PR plots simultaneously which plots performed well. Finally, we will compare CERES and PR plots, and also identify which plot performs better in the detection of heteroscedasticity.

## 2. Material and Methods

In this section, we described CERES and PRplots in detecting outliers in binomial regression. Let us consider the model  $Y = f(X) + \varepsilon$ , (1)

Where  $Y = (y_1, y_2, \dots, y_p)'$  is an  $n \times 1$  vector of response;  $X = (X_1, X_2, \dots, X_p)'$  is a  $n \times 1$  covariate matrix, and  $\varepsilon$  is  $n \times 1$  random vector. The conditional distribution of  $Y$  on  $X$  for a GLM for a set of  $n$  observations due to McCullagh and Nelder (1983) is,

$$d_{y|x}(y|\theta, \psi) = \exp\left\{\frac{\theta y - \mu(\theta)}{v(\psi)} + w(y, \psi)\right\}, \quad (2)$$

where  $\mu(\cdot), v(\cdot), w(\cdot, \cdot)$  are well-known smooth functions;  $\theta$  is an unknown scalar-valued parameter that is dependent on  $X$ , and is  $\psi$  an unknown dispersion parameter.

$$E(Y|X) = \frac{\partial \mu}{\partial \theta} = \mu(x) \text{ and } V(Y|X) = \left\{\frac{\partial^2 \mu}{\partial \theta^2}\right\} v(\psi).$$

There is no consideration of the dispersion parameter  $\psi$ ; when calculating  $\mu(x)$ , as a result,  $v(\psi)$  is presumed to be established. This function's log-likelihood function  $\beta$  is,

$$l(\beta) = \ln L(\beta) = \exp\left\{\frac{\theta y - \mu(\theta)}{v(\psi)} + w(y, \psi)\right\}.$$

The predictors are portioned as  $X' = (X_1', X_2')$ , where  $X_j$  is  $p_j \times 1, j = 1, 2$ . The regression function can be modelled as follows, according to Cook and Croos-Debrera (1998).

$$\eta(x) = h(\mu(x)) = \alpha_0 + \alpha_1' X_1 + g(X_2) \quad (3)$$

Assume that the regression function has a parametric form and that it is given by  $\eta(x) = h(\mu(x)) = \alpha_0 + \alpha_1' X_1 + \alpha_2' X_2$ .

In (3), the term  $h(\mu(x))$  refers to a relation function centered on a monotonic and differentiable probability distribution and  $(\alpha_0 + \alpha_1' + \alpha_2')$  is consisting of a vector of unknown parameters  $(p_1 + 1) \times 1$  vector. The regression function,  $\mu(x) = h^{-1}(\eta(x))$ , is a function of  $x$  or function of  $\eta$  depending on interest and concerns.

The binomial response variable's probability density function is given by

$$f(y; n, \mu) = \binom{n}{y} \mu^y (1 - \mu)^{n-y}, y = 0, 1, 2, \dots, n.$$

It can be written as  $y \sim \text{binomial}(y; n, \mu)$ . The mean and variance of  $y$  are,  $E(y) = n\mu$  and  $\text{var}(y) = n\mu(1 - \mu)$  respectively. In logistic regression, which serves as a binomial example in this article, we begin with a binomial  $(n, \mu)$  random variable  $Y^*|X$ , where the unknown probability of "success"  $p$ , may depend on  $X$ . The known index  $n$  may vary from observation to observation but is assumed to be independent of  $X$ .

The observed fraction of successes from a standard binomial trial is then  $Y = Y^*/n$ . In terms of (2) and (3),

$$\eta = \theta = h(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \quad (4)$$

$\mu(\eta) = \log(1 + \exp(\eta))$ , and  $v(\psi) = 1/n$ . Cook (1993) looked at how well PR plots could depict  $g$  in the special case of additive-error models where the relation is the identity function,  $\eta = \mu$ , and the conditional distribution of  $Y|X$  can be defined as

$$Y|X = \alpha_0 + \alpha_1' X_1 + g(X_2) + \varepsilon, \quad (5)$$

where  $\varepsilon$  is unaffected by  $X$  and has a mean of 0. Cook's research revealed that the output of PR plots is highly influenced by the conditional expectation  $E(X_1|X_2)$ , with the best results obtained when the  $E(X_1|X_2)$ , is linear in the value of  $X_2$ .

Consider summarizing the data by fitting

$$\eta_f(x/b) = h(\mu_f) = b_0 + b_1' X_1 + b_2' \iota(X_2) \quad (6)$$

where  $b' = (b_0, b_1', b_2')$  and  $\iota(X_2)$  is a user-defined  $X_2$  function. The equipped model is indicated by the subscript  $f$  on  $\eta_f$  and  $\mu_f$ . Later, we'll talk about your options for Based on

(6), it is assumed that Estimated coefficients  $\hat{b}_j, j= 0, 1, 2$ , are obtained by minimizing a convex objective function.

$$\hat{b}' = (\hat{b}_0, \hat{b}_1, \hat{b}_2) = \arg \min_b L_N(b), \quad (7)$$

$$L_N(b) = \frac{1}{N} \sum_{i=1}^N L(\eta_f(x_i|b), y_i) = \frac{1}{N} \sum_{i=1}^N L(b_0 + b_1'X_{i1} + b_2'v(X_{i2}), y_i)$$

$L(\cdot, \cdot)$  is a convex objective function with respect to its first argument that is chosen by the consumer. Since it contains ordinary least squares, maximum probability, and some robust estimates, this class is not very restrictive. For logistic regression with the relation provided in (4), for example, the objective function corresponding to maximum likelihood is

$$L(\eta_f(x|b), y) = n\{\log(1 + \exp(\eta_f)) - y\eta_f\}. \quad (8)$$

When we talk about maximum probability, we're talking about the related figures from (2) and (3). (6).

The class of convex objective functions is a generalization of the class of objective functions corresponding to (7).

$$L(\eta_f, y) = L(y - \eta_f)$$

used by Cook (1993) for additive-error models (5). A PR plot for  $X_2$  is obtained by first setting  $v(X_2) = X_2$  and fitting (6) via (7), then constructing the  $(p_2 + 1)$ -dimensional plot  $\{\hat{p}\hat{r}_2, X_2\}$ , where

$$\hat{p}\hat{r}_2 = (y - \hat{u}_f)h'(\hat{u}_f) + \hat{b}'_2X_2 \quad (9)$$

is the partial residual for  $X_2$ ,  $h'(\cdot)$  is the first derivative of  $h(\cdot)$  with respect to  $u$ ,  $\hat{b}$  obtained from (7), and  $\hat{u}_f(x) = h^{-1}(\eta_f(x|\hat{b}))$  is the regression function  $u_f$  evaluated at  $\hat{b}$ . The subscript "2" in  $\hat{p}\hat{r}_2$  is intended to remind that the partial residuals are for  $X_2$ .

Next, to form a CERES plot for  $X_2$  let us set  $v(X_2)$  equal to a function  $E(X_1|X_2)$  that captures the behavior of  $\tilde{E}(X_1|X_2)$ . This function may be  $E(X_1|X_2)$  if known, an estimate  $\hat{E}(X_1|X_2)$  based on smoothing, or a parameterized class of functions that includes  $E(X_1|X_2)$  as a special case. Once  $v(X_2) = \tilde{E}(X_1|X_2)$  is specified, we fit (6) again using (7). The CERES plot for  $X_2$  is then the  $(p_2 + 1)$ -dimensional plot  $\{\hat{c}\hat{r}_2, X_2\}$ , where

$$\hat{c}\hat{r}_2 = (y - \hat{\mu}_f)h'(\hat{\mu}_f) + \hat{b}'_2\tilde{E}(X_1|X_2) \quad (10)$$

is the CERES residual for  $X_2$  constructed from the quantities defined in (8) but based on  $v(X_2) = \tilde{E}(X_1|X_2)$ . A CERES plot reduces to a PR plot when  $\hat{b}'_2\tilde{E}(X_1|X_2)$  is a linear function of  $X_2$ . Cook (1993) provided an extra discussion on the construction of  $\tilde{E}(X_1|X_2)$ .

Partial residuals as defined in (9) reduce to the usual definition of partial residuals in additive-error models (5) because then the link is the identity function and  $h' = 1$ . For logistic regression (4),

$$(y - \hat{\mu}_f)h'(\hat{\mu}_f) = \frac{y - \hat{\mu}_f}{\hat{\mu}_f(1 - \hat{\mu}_f)}$$

and the partial residuals (9) reduce to those defined by Landwehret al. (1984) when the response is binary. Recall that in our formulation,  $y = y^*/n$ . Generally, the first term on the right of (9) can be interpreted in terms of  $\eta$  as the score scaled by the expected information per observation, all evaluated at  $\hat{\mu}_f$ , that  $(y - \mu)h'(\mu) = \frac{\partial \log d_{y|x}/\partial \eta}{-E\{\partial^2 \log d_{y|x}/\partial \eta^2\}}$

Because  $E\left(\frac{\partial \log d_{y|x}}{\partial \eta}\right) = 0$

$$\text{And } -E\left\{\frac{\partial^2 \log d_{y|x}}{\partial \eta^2}\right\} = E\left(\frac{\partial \log d_{y|x}}{\partial \eta}\right)^2$$

$(y - \mu)h'(\mu)$  can also be interpreted as the standardized score weighted by the inverse of its standard deviation. If we let  $\hat{\eta}_f(X) = \hat{\eta}_f(X|\hat{b})$ , the quantity  $\hat{\eta}_f + (y - \hat{u}_f)h'(\hat{u}_f)$  The adjusted dependent variable, which is used in iterative estimation techniques such as the Newton-Raphson process, is often referred to as the adjusted dependent variable (McCullagh and Nelder, 1983). Expression (8) coincides with all partial residual definitions that we are aware of, including those of Collett (1991) and McCullagh and Nelder (1983). However, maximum likelihood estimation is not needed, and 'g' may be a function of multiple predictors, necessitating the use of three-dimensional plots when  $p_2 = 2$ . Fitting a regression curve to the PR plot  $\{\hat{p}\hat{r}_2, X_2\}$  should yield a useful approximation of 'g' up to a linear transformation if the correlation between  $g(X_2)$  and the regression function  $E(\hat{p}\hat{r}_2 | X_2)$  is sufficiently high. Because obtaining a closed-form for  $E(\hat{p}\hat{r}_2 | X_2)$  is difficult.

We use approximation to study the relationship between  $E(\hat{p}\hat{r}_2 | X_2)$  and  $g(X_2)$  and we are going to use  $g(X_2) = bX_2$ .

Consequently, the CERES and PR plots for BRM can be constructed by using equation (9) and (10).

The first derivative of the binomial regression link function given in equation (4) is

$$h'(\hat{\mu}_f) = \frac{1}{\mu(1-\mu)}$$

Hence the fitted model by using log link for binomial regression can be expressed as

$$\hat{\mu}_f = \frac{e^{\hat{\beta}_0 + \hat{\beta}'_1x_1 + \hat{\beta}'_2x_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}'_1x_1 + \hat{\beta}'_2x_2}}$$

Where the regression estimators are  $\hat{\beta}_0, \hat{\beta}'_1, \hat{\beta}'_2$  the fitted model is  $\hat{\mu}_f$  and the predictors are  $x_i$ . Similarly, the CERES and partial residual for a model with p explanatory variables can be expressed as

$$\hat{p}\hat{r}_i = (y - \hat{u}_f)h'(\hat{u}_f) + \hat{b}'_iX_i \quad i = 1, 2, \dots, p. \quad (11)$$

$$\hat{c}\hat{r}_i = (y - \hat{\mu}_f)h'(\hat{\mu}_f) + \hat{b}'_i\tilde{E}(X_i|X_i) \quad i = 1, 2, \dots, p. \quad (12)$$

In addition, for explanatory variables, the fitted model is

$$\hat{\mu}_f = \frac{e^{\hat{\beta}_0 + \hat{\beta}'_1x_1 + \hat{\beta}'_2x_2 + \dots + \hat{\beta}'_px_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}'_1x_1 + \hat{\beta}'_2x_2 + \dots + \hat{\beta}'_px_i}} \quad (13)$$

In the next section, the real data is used to elaborate the theoretical part of CERES and PR plots using liver cancer data.

### 3. Example Liver Cancer data

Areal data (see appendix ) would be used to assess the performance of CERES and PR plots in the detection of heteroscedasticity. The methodology developed in the previous section is implemented on the Liver cancer data that was utilized by Zelterman (1999) and also later by Atkinson and Riani (2001). The response variable ( $Y$ ) that follows a binomial distribution with two explanatory variables which are, dose of a patient ( $X_1$ ), and months of study ( $X_2$ ) that contains 72 observations. In order to detect heteroscedasticity, the CERES and PR plots for the Liver cancer data are shown

in Figures 1 and 2, respectively. Since, we have two predictors in the model therefore there are two possible CERES and PR plots that would be obtained.

The summary of binomial regression model for Liver cancer data is presented in Table 1. Based on the result, it shows that both of independents variables are significant ( $p$ -value < 0.05).

Table 1. Binomial Regression Analysis for Liver Cancer Data

Predictors	Coefficients	Standard error	t-test	$p$ -value
Constant	0.411	0.124	3.32	0.001
$X_1$	0.1972	0.0905	2.18	0.033

$X_2$	0.01788	0.00588	3.04	0.003
-------	---------	---------	------	-------

$$R^2 = 16.88\%, R^2(adj) = 14.47\%$$

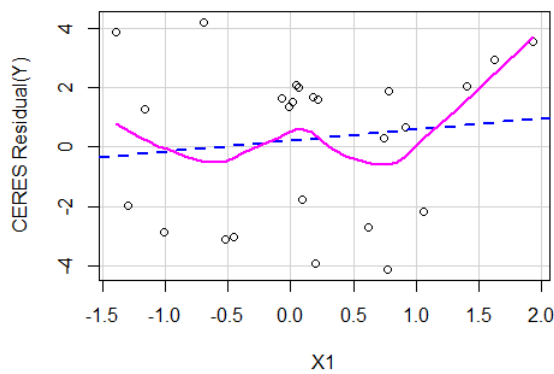
$$\hat{Y} = 0.411 + 0.1972 X_1 + 0.01788 X_2$$

To check the heteroscedasticity in Liver cancer data, we applied Levene's test. It is also observed that the test also has a significant  $p$ -value which raised the problem of heteroscedasticity in the dataset (Table 2).

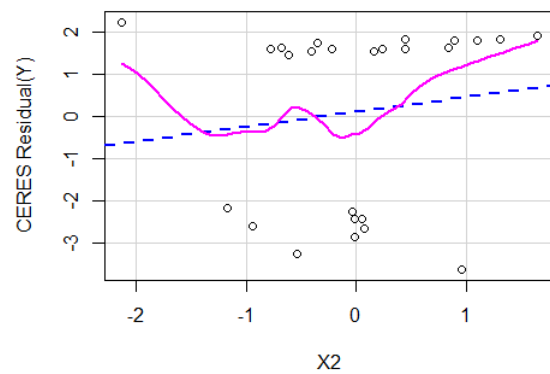
Table 2. Regression diagnostic test

Test Statistic	Statistic	$p$ -value
Levene's test	52.072	0.0000

CERES Plots



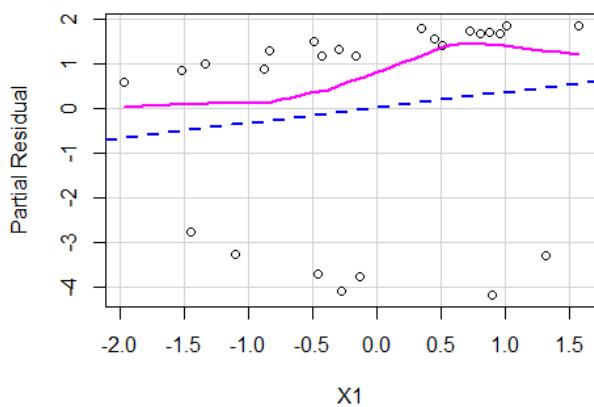
(a). CERES Plot ( $X_1$  = Dose of a patient) for Heteroscedasticity



(b). CERES Plot ( $X_2$  = Month of study) for Heteroscedasticity

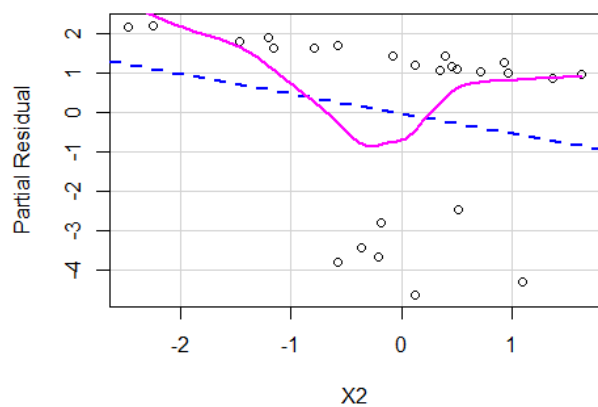
Figure 1: CERES plots for binomial regression model for Liver Cancer Data

Partial Residual Plots



(c). PR Plot ( $X_1$  = Dose of a patient) for Heteroscedasticity

Figure 2: Partial residual plots for binomial regression model for Liver Cancer Data



(d). PRPlot ( $X_2$  = Month of study) for Heteroscedasticity

from the trend line. However, PR plots detect heteroscedasticity more better as compared to CERES plots as due to the larger disparity. Based on the effectiveness of these plots, we can propose that both of these plots can be used as another alternative method to detect heteroscedasticity in addition to the formal statistical tests. In the next section, we also tried to detect heteroscedasticity by using a simulated dataset.

## 4. Simulation Study

In this study, we follow the Monte Carlo simulation for heteroscedasticity used by Cribari-Neto (2004). Aslam et al. (2013) also have followed that scheme. The numerical scheme and the relevant model for this simulation is given as.

$$X_{ij} = \sqrt{(1 - \theta^2)}Z_{ij} + \theta Z_{i(j+1)} \quad i=1,2, \dots, n, j=1, 2, \dots, p$$

Where  $Z_{ij}$  is generated by the standard normal distribution i.e.  $Z_{ij} \sim N(0,1)$  and  $\theta$  is the level of multicollinearity set as 0.8, 0.9, 0.95, and 0.99 in the above simulation equation.

$$\hat{\mu}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}'_1 x_1 + \hat{\beta}'_2 x_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}'_1 x_1 + \hat{\beta}'_2 x_2}}$$

The response variable is generated randomly as  $y \sim B(1, \hat{\mu}_i)$ .

The regression coefficients are considered to be fixed as  $\beta_0 = \beta_1 = \beta_2 = 1$ .

The distribution of error term,  $u_i = \sigma_i \varepsilon_i$  is normal with zero mean and standard deviation  $\sigma_i$ . Moreover, it is assumed the  $\varepsilon_i$ 's to be observed independent. The variance of error terms generated as;

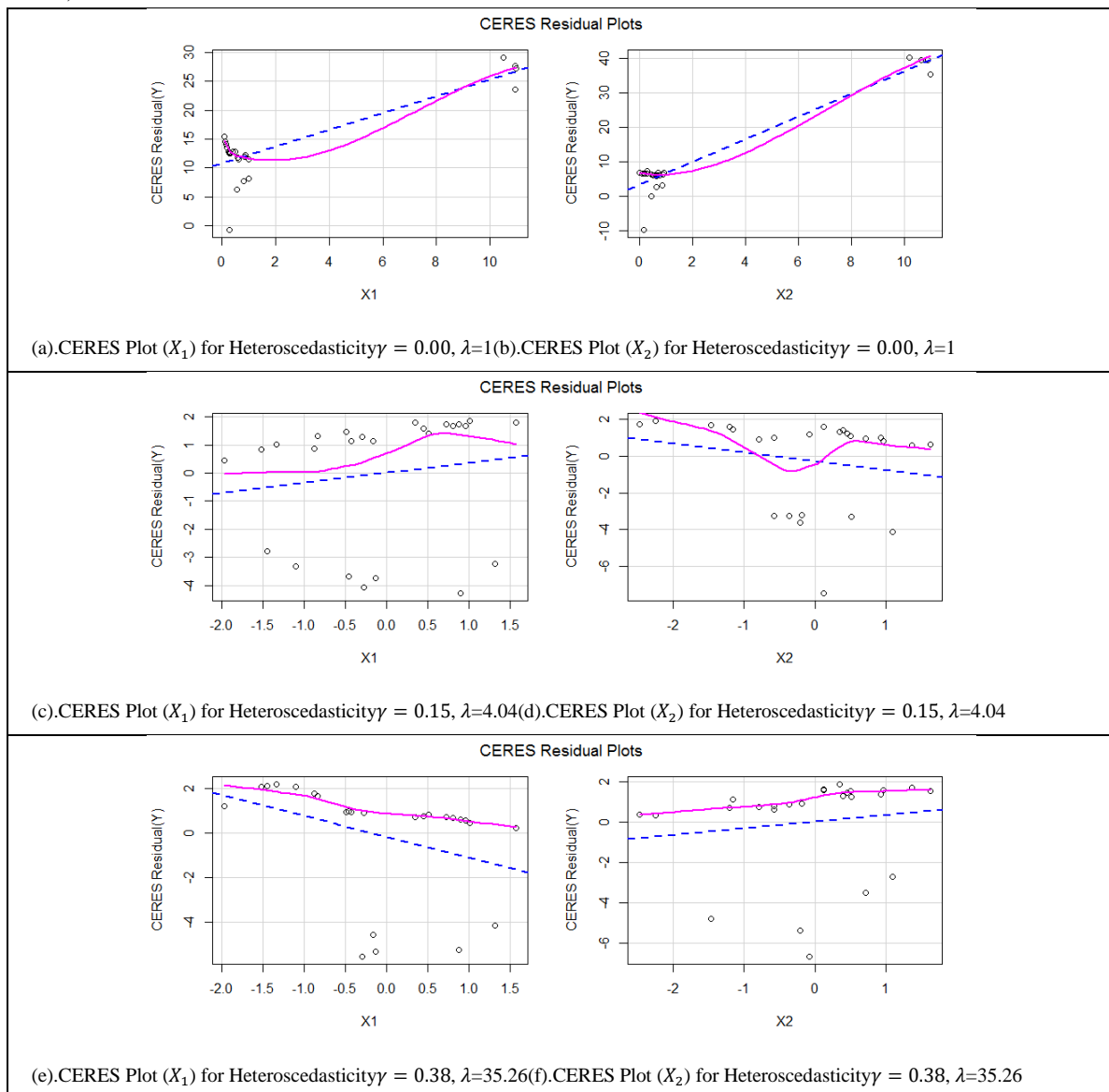
$$\sigma_i^2 = \exp\{\gamma X_{i1} + \gamma X_{i2}^2\}$$

where the values of  $\gamma$  vary as follow;  $\gamma = 0.00, 0.15, 0.38, 0.75$ . The measure in extent to the degree of heteroscedasticity is determined by,

$$\lambda = \frac{\max(\sigma_i^2)}{\min(\sigma_i^2)}$$

$\lambda = 1$  for the case of homoscedasticity,  $\lambda > 1$  for heteroscedasticity.

We have selected four different sample sizes, i.e.,  $n$  are selected as 25, 50, 100, and 200. Each of the result is based on 10,000 simulations. The performance of the CERES and PR plots and diagnostics are assessed. The simulation is conducted using the R software. The graphical displays of the CERES and the PR plots are presented in Figures 3 to 10.



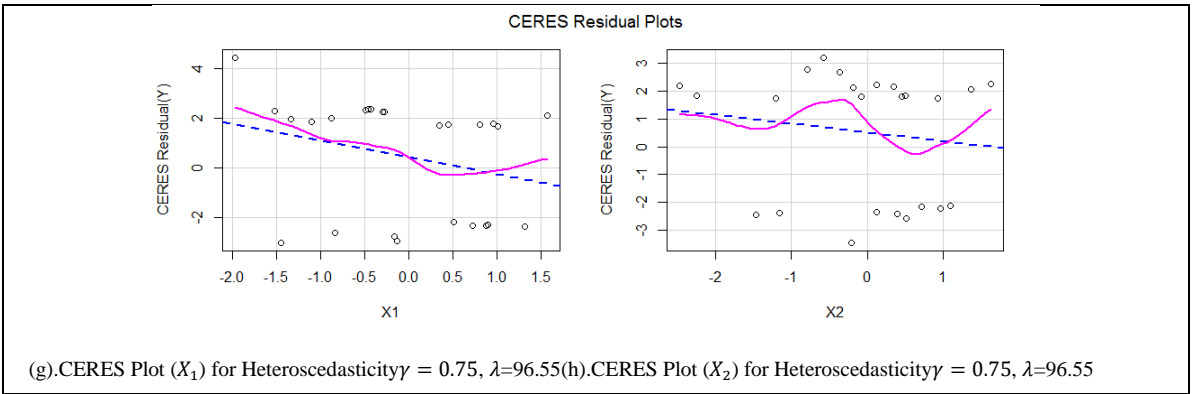
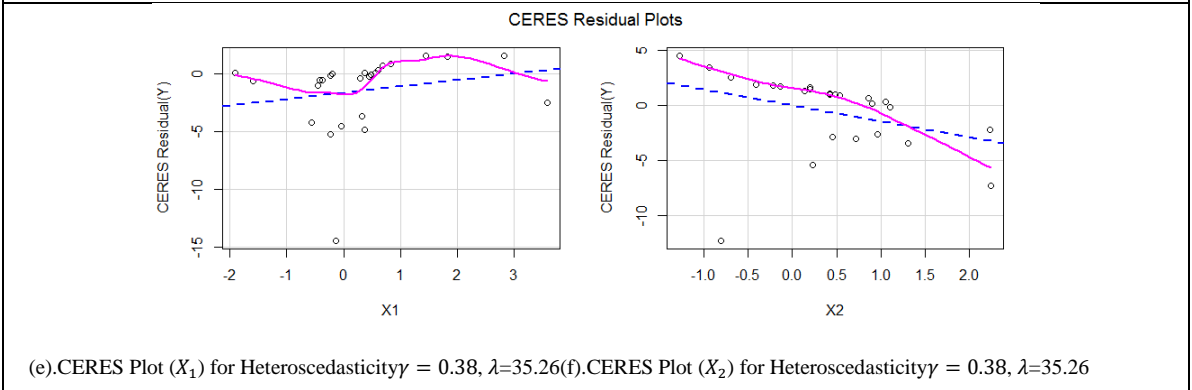
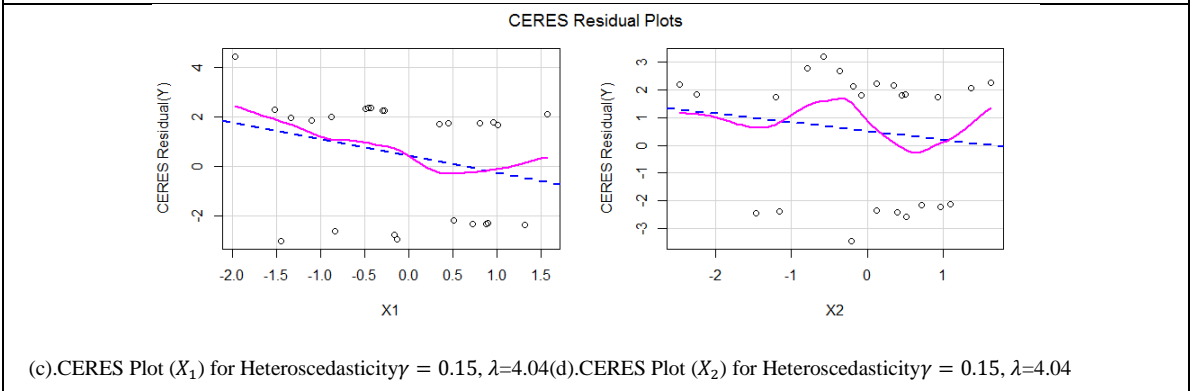
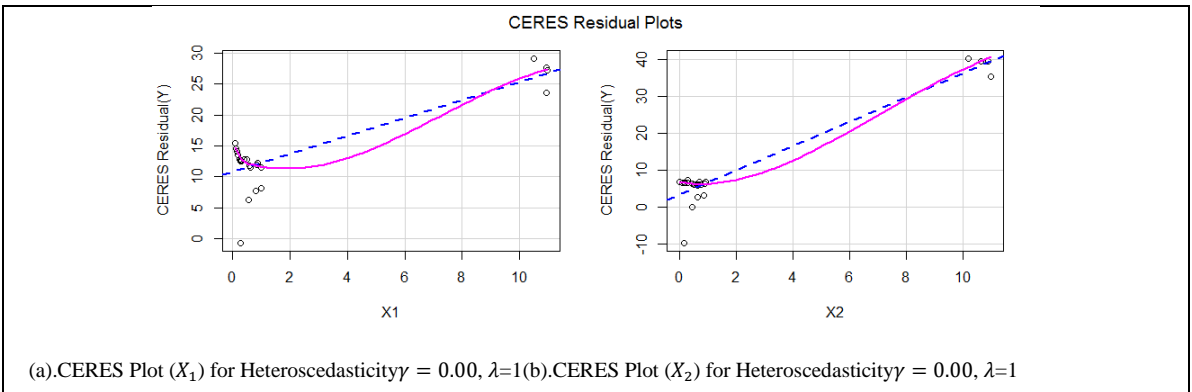


Figure 3: CERES plots for binomial regression model for simulated data, when  $n=25$



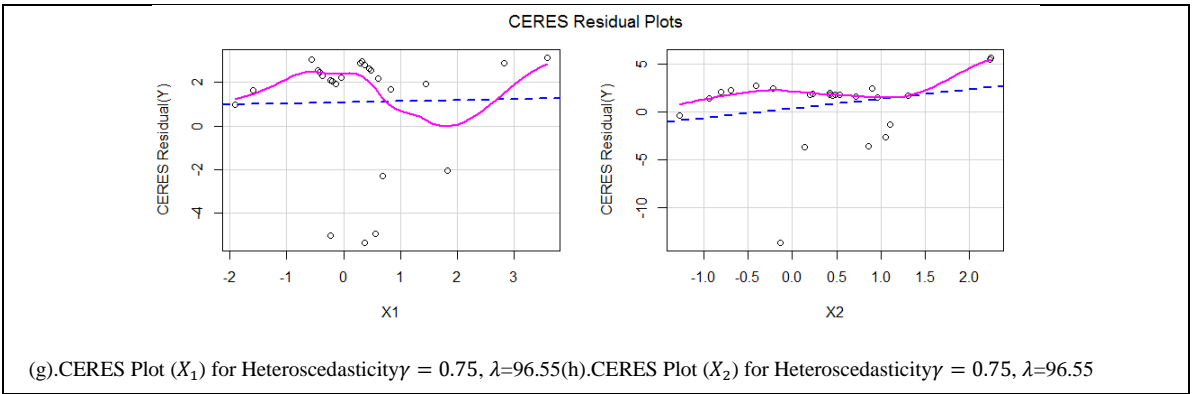
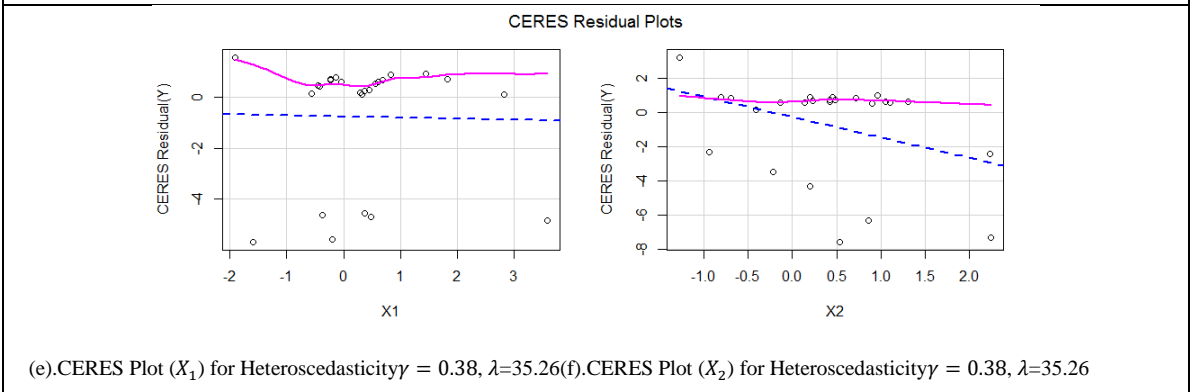
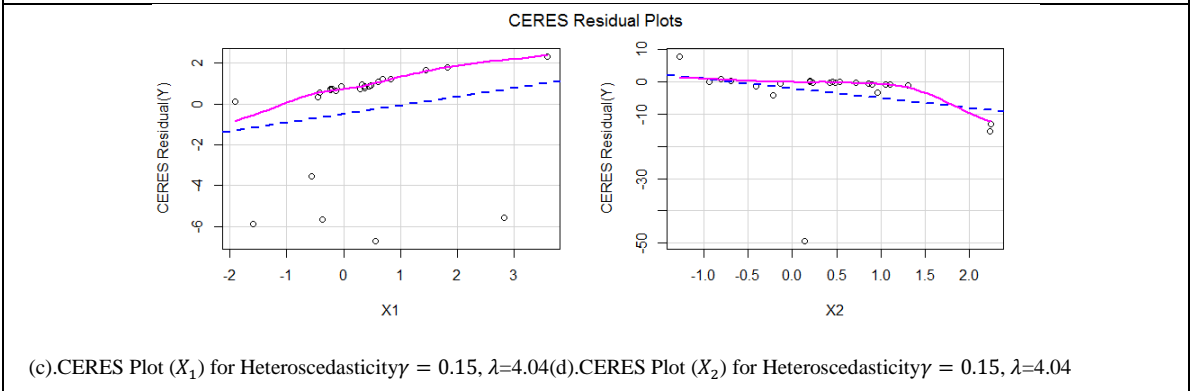
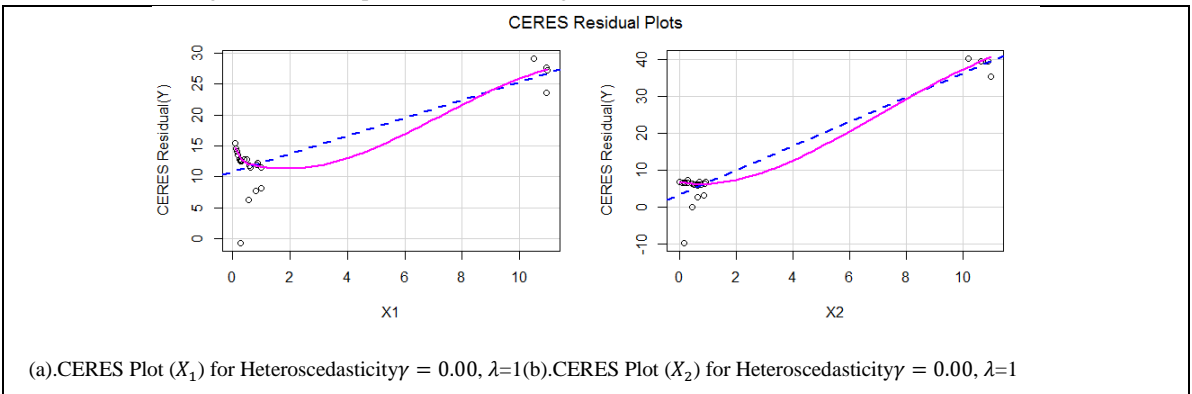


Figure 4: CERES plots for binomial regression model for simulated data, when  $n=50$



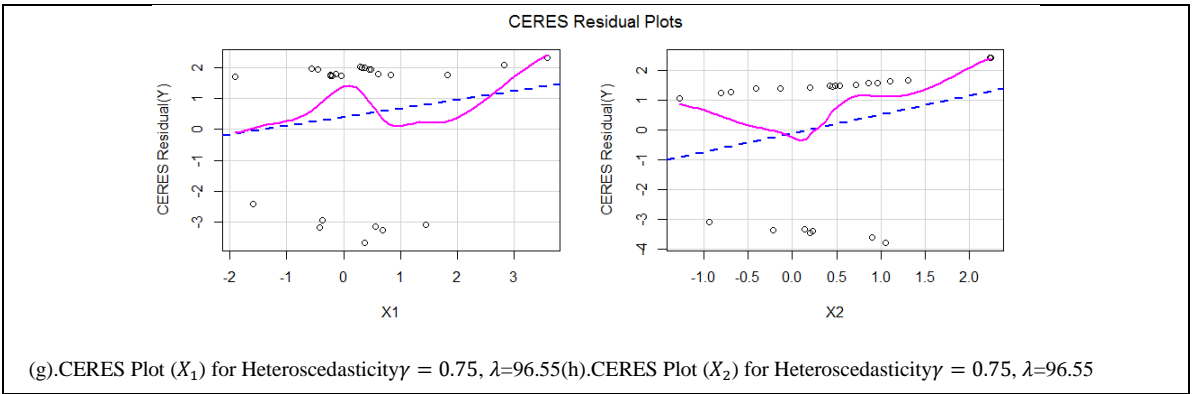
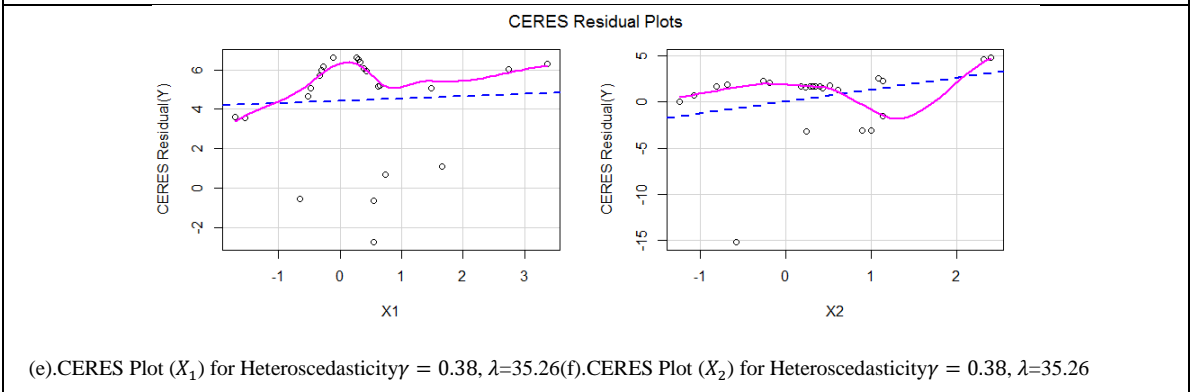
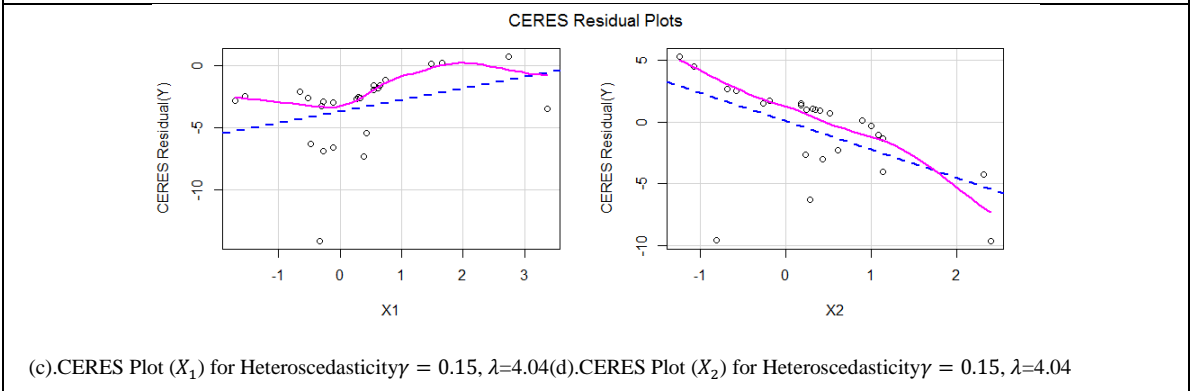
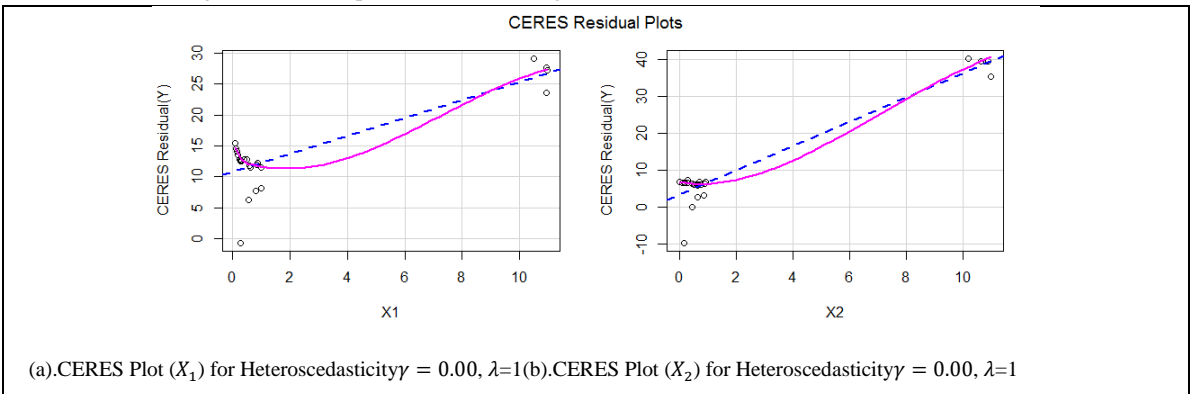


Figure 5: CERES plots for binomial regression model for simulated data, when  $n=100$





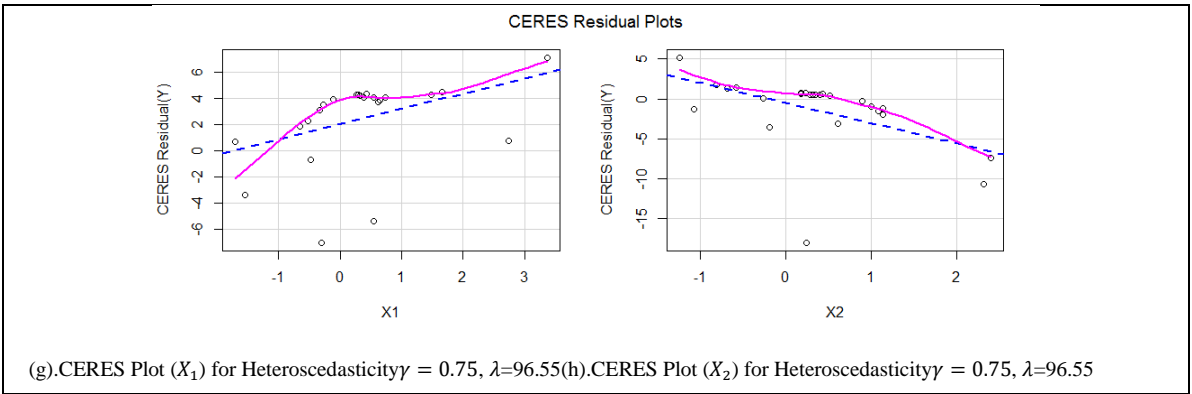
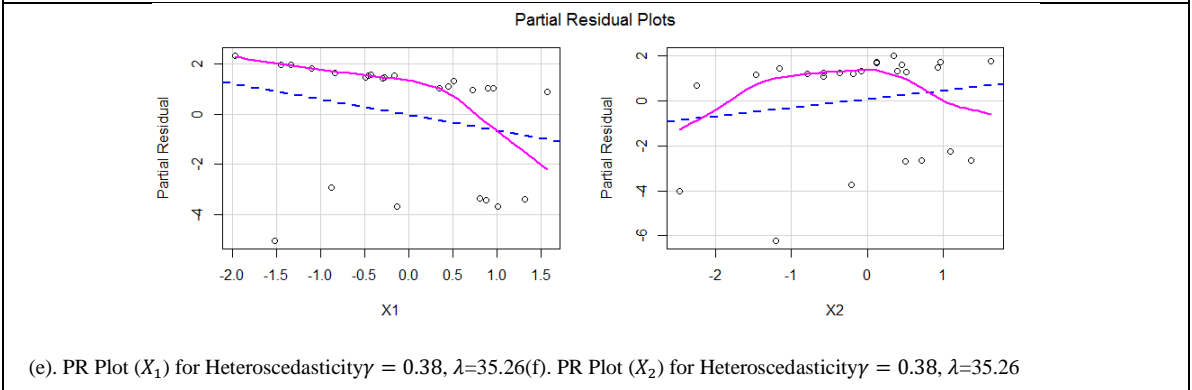
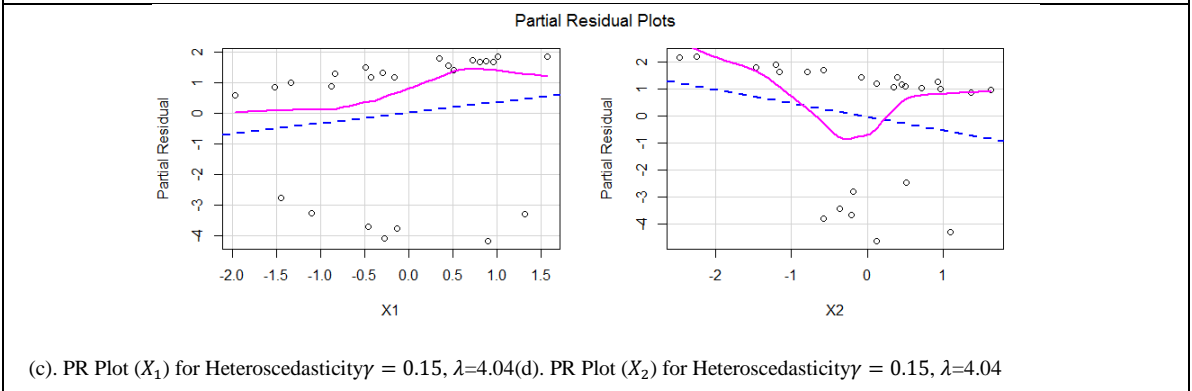
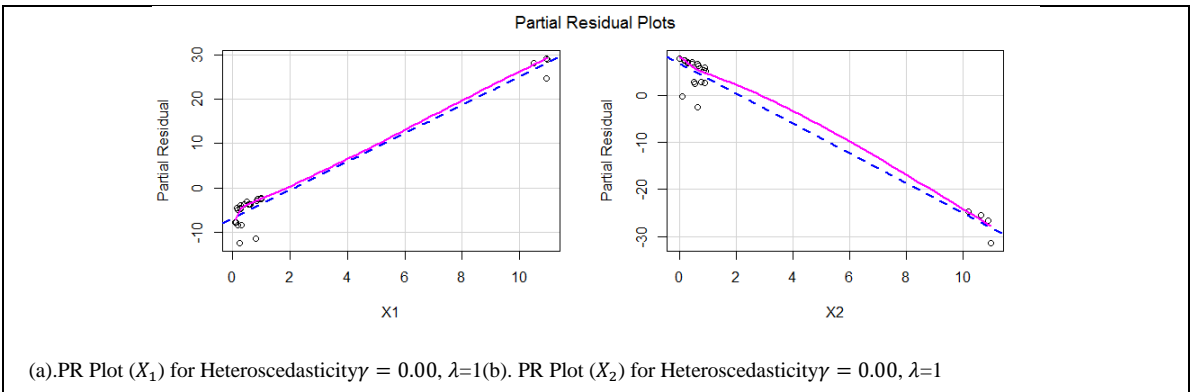


Figure 6: CERES plots for binomial regression model for simulated data, when  $n=200$



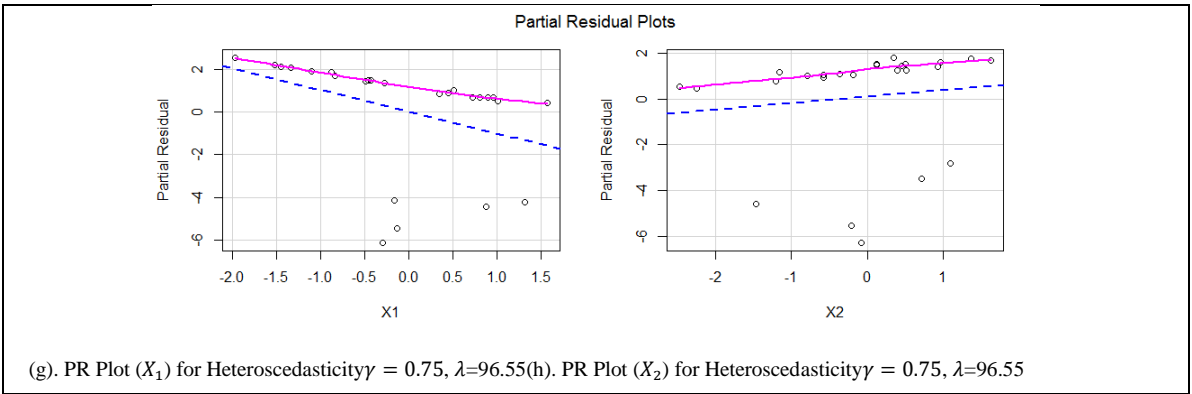
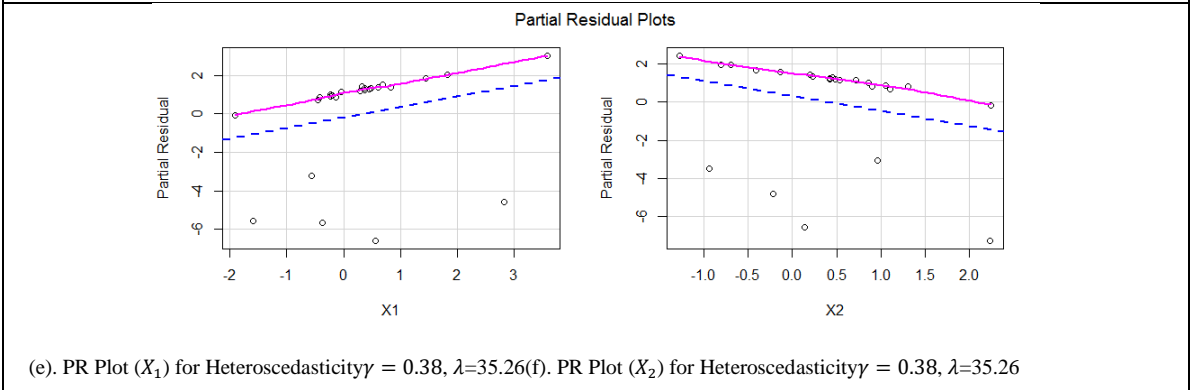
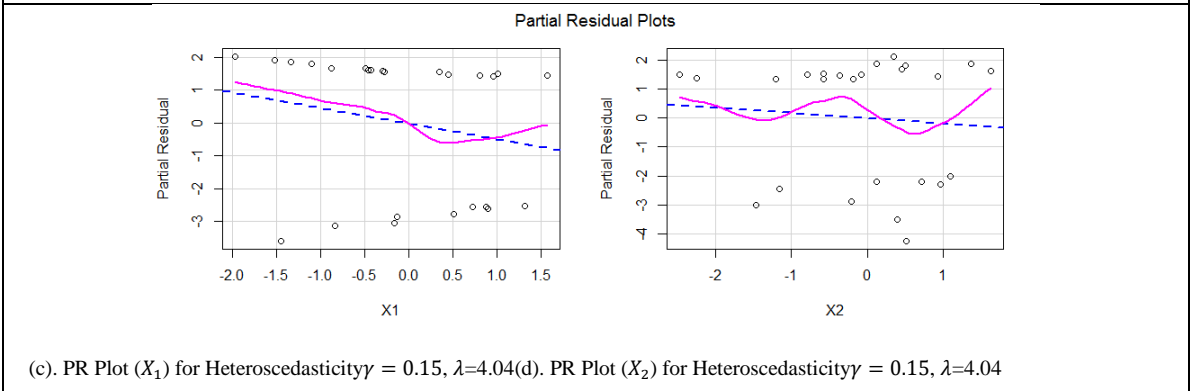
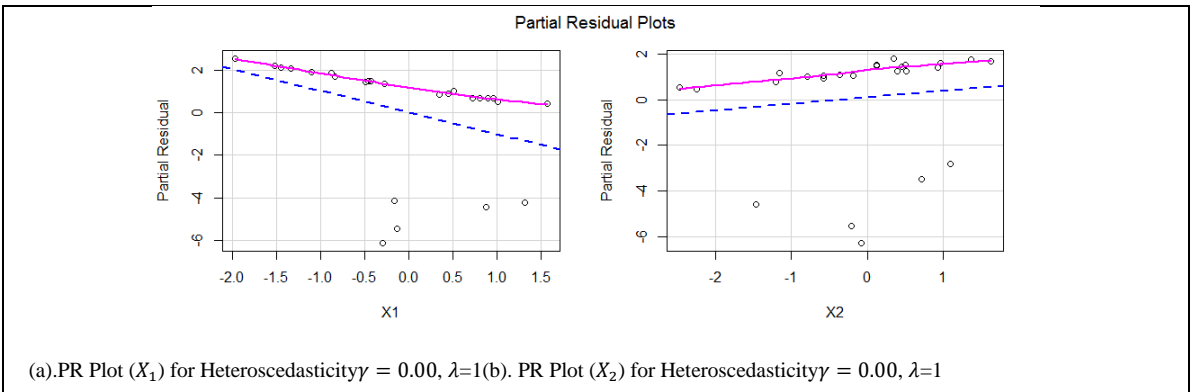


Figure 7: Partial residual plots for binomial regression model for simulated data, when  $n=25$



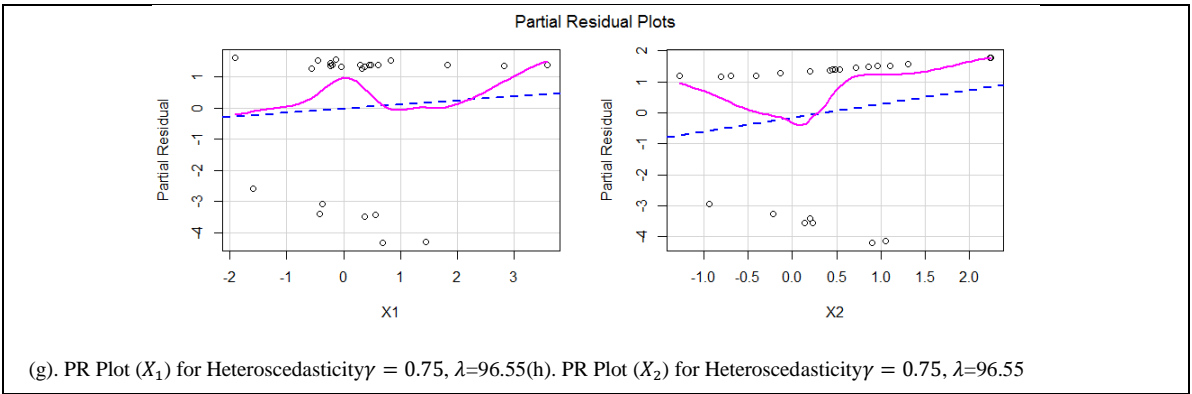
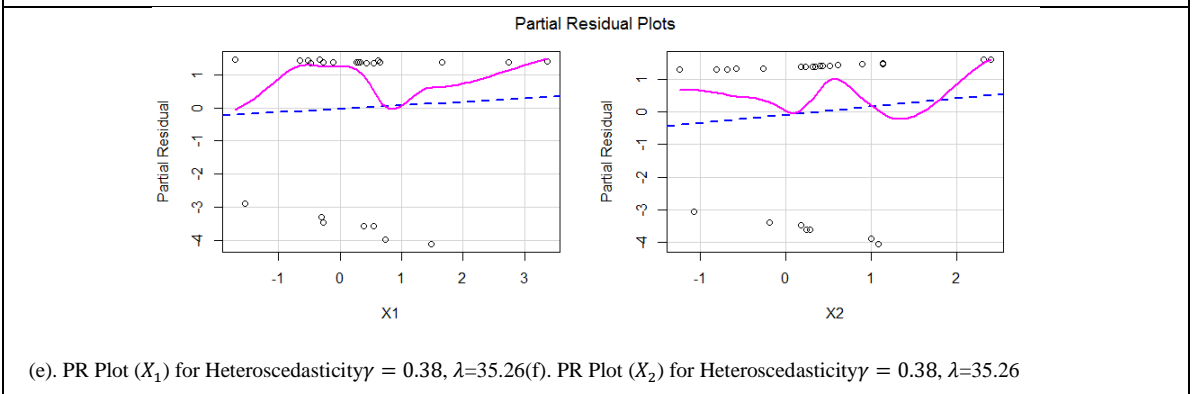
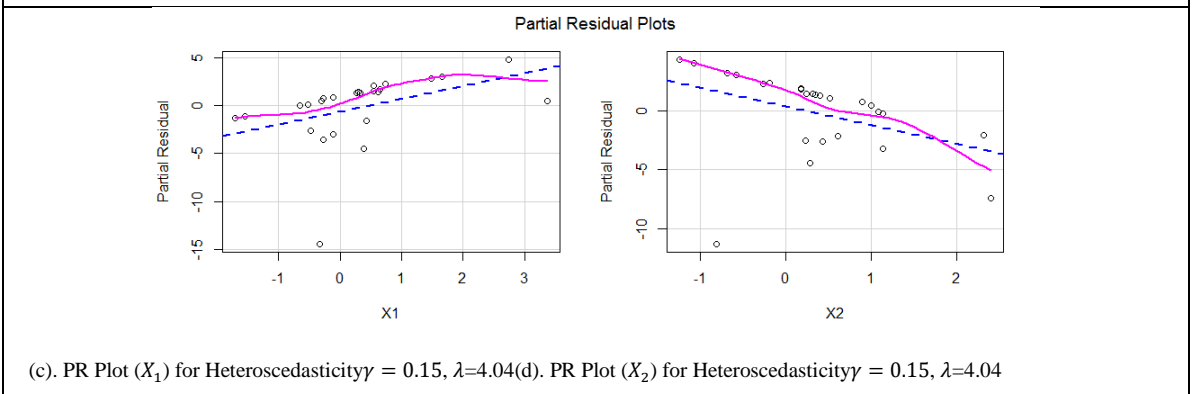
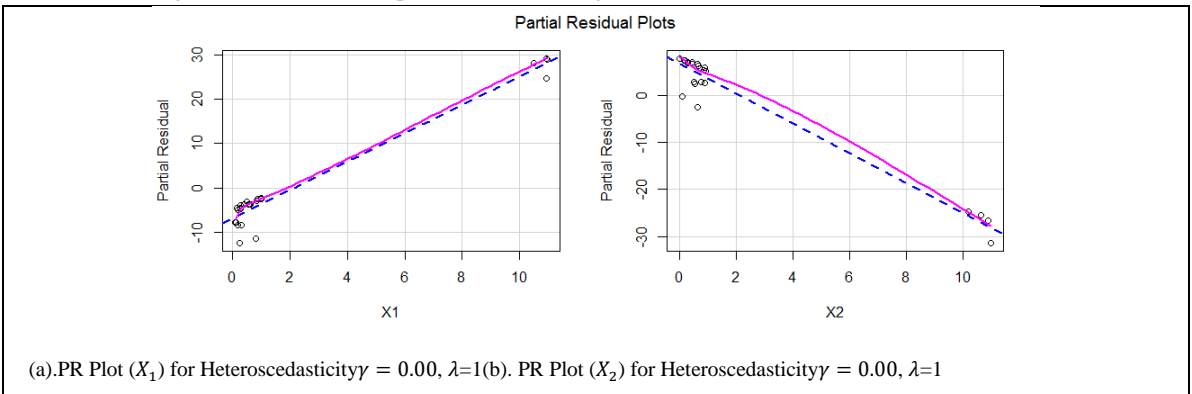


Figure 8: Partial residual plots for binomial regression model for simulated data, when  $n=50$



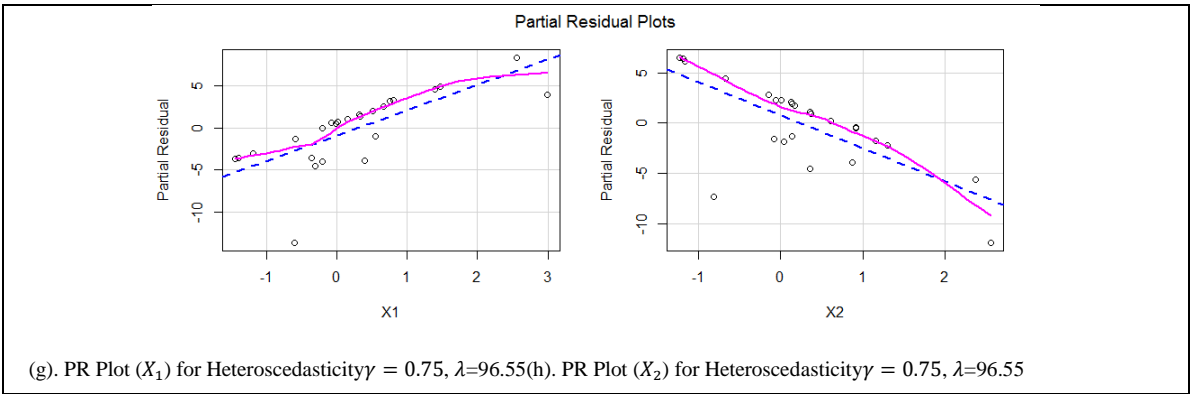
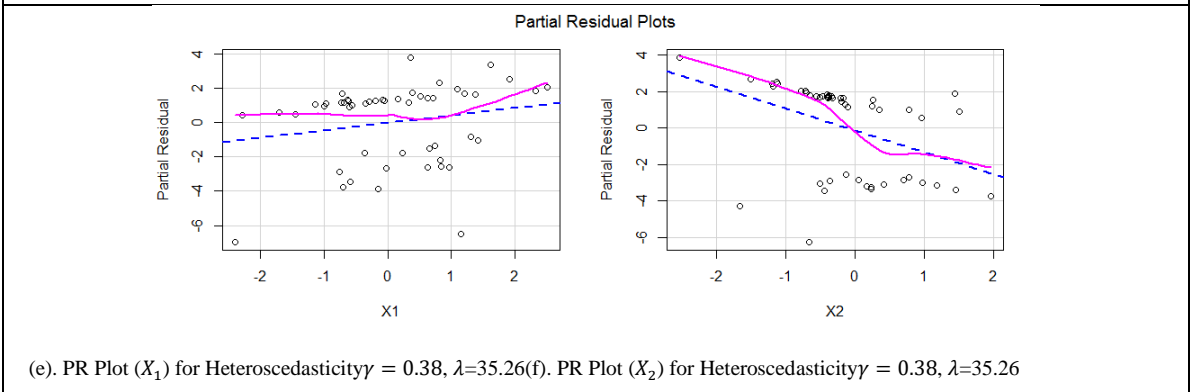
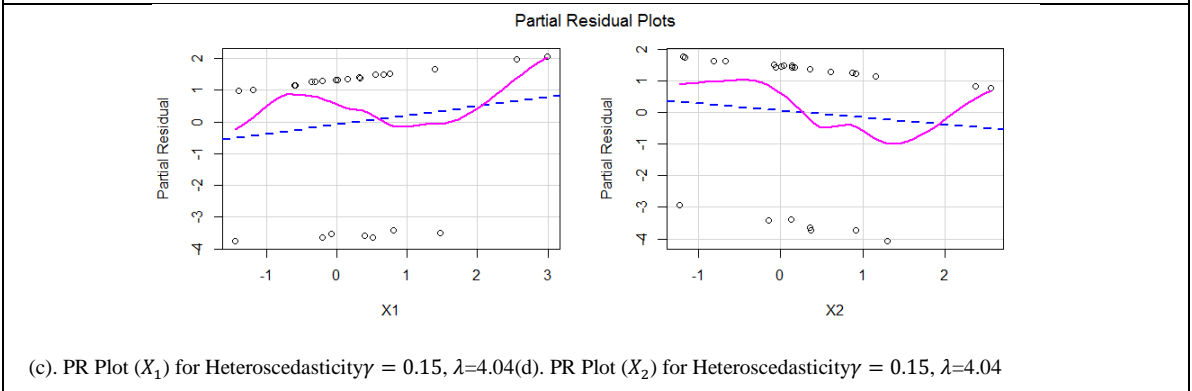
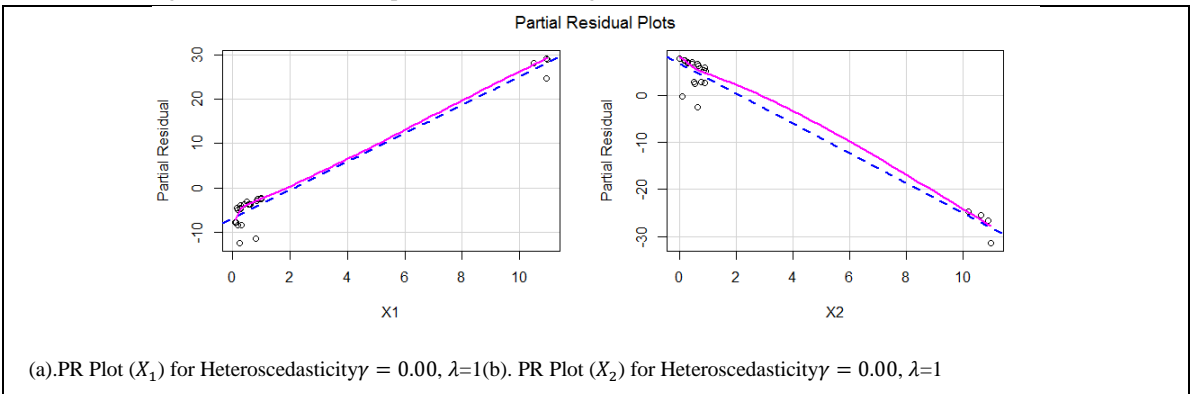


Figure 9: Partial residual plots for binomial regression model for simulated data, when  $n=100$



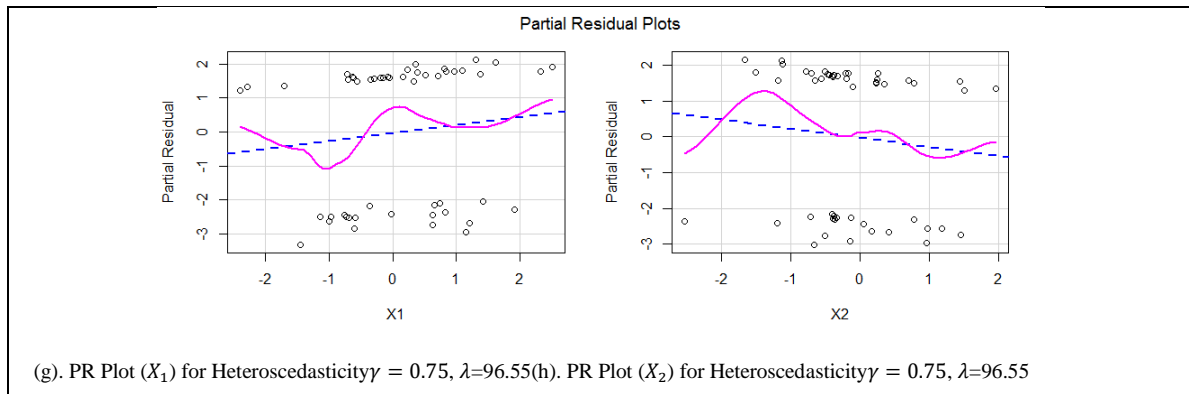


Figure 10: Partial residual plots for binomial regression model for simulated data, when n=200

By using the simulated data set, heteroscedasticity was clearly detected in Figures 3 to 6 present the CERES plots while Figures 7 to 10 present the PR plots. Firstly, we used here, sample size of n=25,50,100,200 observations in simulated data. The CERES and PR plots for heteroscedasticity when sample size n = 25,  $\gamma = 0.00, 0.15, 0.38, 0.75$  and  $\lambda = 1, 4.04, 35.26, 96.55$  if  $\lambda = 1$  it is a case of homoscedasticity and if  $\lambda > 1$  it is a case of heteroscedasticity. Similarly, we each sample size n = 50, 100, 200 for the heteroscedasticity. It is observed that CERES and PR plots detect heteroscedasticity for regressor ( $x_1, x_2$ ). Because various observations are far from the trend line. However, in overall PR plots detect heteroscedasticity more better as compared to CERES plots as due the larger disperity. In other words, both plots showed more observations are far away to each other's that is dispersed between the points. The heteroscedasticity in the PR plot shows more dispersed between the points as compare to the CERES plot. The PR plots gives better visual diagnostic for heteroscedasticity as compare to CERES plots.

## 5. Conclusions

This article addresses the development and implementation of CERES and PR plots for the identification of heteroscedasticity in a binomial regression model. At first, we develop a methodology and then apply it to real-life and simulated data. Both real and simulated data reveal that the proposed method can successfully detect the heteroscedasticity problem in BRM. Both CERES and PR plots perform well in doing this job. But the PR plot performs well better than the CERES plot in the detection of heteroscedasticity.

## Appendix

A<sub>1</sub>: Liver cancer data

Observations	Total number tasted	Number with Cancer	Dose of a patient	Months on Study
1	199	0	0.00	9
2	147	1	0.30	9
3	76	1	0.35	9
4	52	0	0.45	9
5	345	0	0.60	9

6	186	0	0.75	9
7	168	1	1.00	9
8	169	1	1.50	9
9	164	0	0.00	12
10	151	1	0.30	12
11	27	1	0.35	12
12	14	1	0.45	12
13	283	1	0.60	12
14	153	0	0.75	12
15	149	1	1.00	12
16	152	1	1.50	12
17	133	1	0.00	14
18	42	1	0.30	14
19	25	0	0.35	14
20	14	1	0.45	14
21	243	1	0.60	14
22	124	0	0.75	14
23	127	1	1.00	14
24	127	1	1.50	14
25	115	0	0.00	15
26	75	1	0.30	15
27	35	1	0.35	15
28	20	0	0.45	15
29	203	1	0.60	15
30	109	1	0.75	15
31	99	1	1.00	15
32	100	1	1.50	15
33	205	1	0.00	16
34	66	1	0.30	16
35	61	1	0.35	16
36	304	1	0.45	16
37	287	1	0.60	16
38	193	1	0.75	16
39	100	1	1.00	16
40	110	1	1.50	16

41	153	0	0.00	17
42	69	1	0.30	17
43	443	1	0.35	17
44	302	1	0.45	17
45	230	1	0.60	17
46	166	1	0.75	17
47	85	1	1.00	17
48	82	1	1.50	17
49	555	1	0.00	18
50	2014	1	0.30	18
51	1102	1	0.35	18
52	550	1	0.45	18
53	411	1	0.60	18
54	382	1	0.75	18
55	213	1	1.00	18
56	211	1	1.50	18
57	762	1	0.00	24
58	2109	1	0.30	24
59	1361	1	0.35	24
60	888	1	0.45	24
61	758	1	0.60	24
62	587	1	0.75	24
63	297	1	1.00	24
64	314	1	1.50	24
65	100	1	0.00	33
66	445	1	0.30	33
67	100	1	0.35	33
68	103	1	0.45	33
69	67	1	0.60	33
70	75	1	0.75	33
71	31	1	1.00	33
72	11	1	1.50	33

Source: Atkinson, A. C., & Riani, M. (2001). Regression diagnostics for binomial data from the forward search. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(1), 63-78.

## Reference

- Aslam, M., Riaz, T., & Altaf, S. (2013). Efficient estimation and robust inference of linear regression models in the presence of heteroscedastic errors and high leverage points. *Communications in Statistics-Simulation and Computation*, 42(10), 2223-2238.
- Atkinson, A. C., & Riani, M. (2001). Regression diagnostics for binomial data from the forward search. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(1), 63-78.
- Berk, K. N., & Booth, D. E. (1995). Seeing a curve in multiple regression. *Technometrics*, 37(4), 385-398.
- Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlations (with discussion). *Journal of the American Statistical Association*. 80 (391), 580-619.
- Collett, D. (1991). Modeling Binary data, London: Chapman and Hall.
- Cook, R. D. (1993). Exploring partial residual plots. *Technometrics*, 35(4), 351-362.
- Cook, R. D., Croos-Dabrera, R., (1998). Partial residual plots in generalized linear models. *Journal of the American Statistical Association*, 93(442), 730-739.
- Ezekiel, M. (1924). A method of handling curvilinear correlation for any number of variable. *Journal of the American Statistical Association*, 19(148), 431-453.
- Fowlkes, E. B. (1987), Somediagnosics for binary logistic regression via smoothing. *Biometrika*, 74, 503-515.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.
- Imran, M., & Akbar, A. (2020). Diagnostics via partial residual plots in inverse Gaussian regression. *Journal of Chemometrics*, 34(1), e3203.
- Kahng, M. W., & Lee, E. J. (2004). CERES plot in generalized linear models. *Communications for Statistical Applications and Methods*, 11(3), 575-582.
- Landwehr, J. (1983). Using partial residual plots to detect nonlinearity in multiple regression. *Unpublished manuscript. Bell Laboratories, Murray Hill, New Jersey.*
- Landwehr, J. M., and Pregibon, D. (1993), Comments on 'Improved added variable and partial residual plots for the detection of influential observations in generalized linear model' by R. J. O' Hara Hines and E. M. Carter, *Applied Statistics*, 42, 16-19.
- Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984), Graphical methods for assessing logistic regression models, *Journal of the American Statistical Association*, 79, 61-83.
- Larsen, W. A., & McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, 14(3), 781-790.
- Liu, L., Chen, X., Liu, W., Yu, H., & Liu, F. (2019). Statistical analysis and heuristic identification of unexpected interactions from the neurokinase-inhibitor interactome in trigeminal neuralgia pharmacological intervention. *Journal of Chemometrics*, 33(6), e3126.
- Lukman, A. F., Ayinde, K., Binuomote, S., & Clement, O. A. (2019). Modified ridge-type estimator to combat multicollinearity: Application to chemical data. *Journal of Chemometrics*, 33(5), e3125

- 
19. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.
  20. Mallows, C. L. (1986). Augmented partial residuals. *Technometrics*, 28(4), 313-319.
  21. Mansfield, E. R., & Conerly, M. D. (1987). Diagnostic value of residual and partial residual plots. *American Statistician*, 41(2), 107-116.
  22. McCullagh, P. & Nelder, J. A. (1983). *Generalized Linear Models*. London: Chapman and Hall.
  23. Mullet, G.M. (1976), Why regression coefficients have the wrong sign, *Journal of Quality Technology*. 8, 121-126.
  24. Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A* 135, 370-84.
  25. Özkale, M. R., Lemeshow, S., & Sturdivant, R. (2018). Logistic regression diagnostics in ridge regression. *Computational Statistics*, 33(2), 563-593.
  26. Park, M. Y., & Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4), 659-677.
  27. Wood, F. S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics*, 15(4), 677-695.
  28. Wouters, J. M., Gusmao, J. B., Mattos, G., & Lana, P. (2018). Polychaete functional diversity in shallow habitats: Shelter from the storm. *Journal of Sea Research*, 135, 18-30.
  29. Zelterman, D. (1999) *Models for Discrete Data*. Oxford university Press.