https://dx.doi.org/10.4314/dujopas.v8i1a.15
ISSN (Print): 2476-8316
ISSN (Online): 2635-3490
Dutse Journal of Pure and Applied Sciences (DUJOPAS), *Vol. 8 No. 1a March 2022*

Estimation of Time Independent Cox Proportional Hazard Model with Correlated Unordered Categorical Covariates

Onatunji Adewale Paul[1], Olalude Oladapo Ayokomi [2] and Oladimeji Luqman Abiola[3]

[1,3] Department of Statistics,
Ladoke Akintola University of Technology,
Ogbomoso, Oyo State, Nigeria.

[2] Department of Statistics,
University of Ibadan,
Oyo State, Nigeria.

E mail: aponatunji@lautech.edu.ng

Abstract
Many studies have been carried out since the development of Cox model(1972). Little or no attention on the performance of Cox model with correlated unordered categorical data exists in literature. The paper is aimed at investigating the behavior of the parameter estimates and the model under varied collinearity (ρ =0.2, 0.6, 0.9) between discrete independent variablse. Two discrete independent data were generated using binomial distribution with the considered levels of collinearity for sample size of 10, 30, 100 and 500.Survival time was generated with estimated baseline hazard function. The paper shows that regression coefficients with small standard errors are statistically significant when sample size is large at low collinearity as against small sample size. There is evidence of non proportional hazard assumptions when there exists collinearity between unordered categorical data. Moreso, Cox model is statistically significant at all levels of collinearity considered for varied sample size except when sample size is 30 in the study.

Keywords: cox proportional hazard model

INTRODUCTION
In survival analysis, collinearity occurs either by structural function or data in which one predictor variable can be linearly explained by other predictor variables. Violation of basic assumption, collinearity in proportional hazard model causes poor interpretation of the estimates as a result of biased estimates. Proportional hazard model developed by Cox(1972) is frequently used in survival analysis to assess time to event response variable, and for evaluating the effects of predictor variables despite the violation of proportionality of hazard. Babalola and Yahya (2019) studied the effects of collinearity on the time dependent coefficients of Cox model and found that collinearity influenced the correctness of the estimates of the covariates in the framework of this model. Collinearity among the covariates in survival studies causes unstable estimates and large variance. Thus, effects of important collinear covariates are wrongly interpreted. In order to reduce effect of the collinearity,

**Author for Correspondence*

Lagakos(1988) categorized those correlated continuous covariates but reported the nontrivial loss of efficiency in the results. Biased estimates were also observed when the variables were removed. The study of censored survival phenotypes is more informative than treating the phenotypes as categorical variables in survival analysis due to large variability in time to certain clinical event among patients. This method of regression is useful in the areas of health and environmental study because of its flexibility especially when there is no need of distributional assumptions. Method for fitting the proportional hazards regression model when the data are interval-censored observations to test the hypothesis of a zero regression coefficient that led to a generalization of the log-rank test for comparison of several survival curves(Finkelstein, 1986). Examples on application of the cox model are duration of birth process(Triastuti *et al.*, 2018), consumer purchase decision on product(Azimmatul, 2014), assessment of health risks associated with occupational radiation exposure(Xue *et al.*, 2007). Ledwon and Jäger (2020) developed and applied Cox proportional hazards regression for time-dependent covariates to assess corporate insolvency for non-financial constitutes represented in CDAX. CPH models considered Andersen-Gill counting process (AG-CP) to explore the importance of accounting and financial ratios and industry effects that are useful in detecting potential insolvencies. Finings revealed that industry grouping added marginal predictive power and no overall improvement in accuracy rates when market variables are already included in the probability of default (PD) model. This study is in contrary to the findings of Chava and Jarrow (2004). There are many studies on cox proportional hazard model with passage of time in the literature. However, little or no work has been done on possibility of having collinearity between two categorical data. This collinearity violates the Cox model assumption; and it cannot not only cause biased estimates but also misleading interpretation.

The goal of this study is to examine the performance of Cox model when covariates that are correlated with non-proportionality are unordered categorical data. The organization of the paper is as follows: methodology is considered in section 2, section 3 gives brief description of simulation study, section 4 presents discussion of results and section 5 discusses conclusion.

METHODOLOGY

Cox model is a semiparametric model when baseline hazard function is unspecified and that covariates enter the model linearly. Then, Cox(1972) proposed the model below

$$h_i(t|X_i') = \exp(\beta_i x_{i1}(t) + \beta_2 x_{i2}(t) + ... + \beta_k x_{ik}(t)) \quad (1)$$

$X' = (x_{i1}, ..., x_{ik})$ is a vector of unordered categorical explanatory variables and $\beta' = (\beta_{i1}, ..., \beta_{ik})$ the fixed time regression coefficient. Equation (1) multiplied by baseline hazard to become cox proportional hazard model given in equation(2), this is a semiparametric model as a result of the presence of baseline hazard, $h_0(t) > 0$ not specified. $X_i' = T(X_i)$ is a binary indicator of treatment group measured 0 and 1. Consider two observations i and j, for collinear unordered categorical(Binary Independent) variables, x *values* without intercept term, that linearly enter the model.

$$h_i(t|X_i) = h_0(t)\exp(\beta_i x_{i1}(t) + \beta_2 x_{i2}(t) + ... + \beta_k x_{ik}(t)) \quad (2)$$

and

$$h_j(t|X_j) = h_0(t)\exp(\beta_j x_{j1}(t) + \beta_2 x_{j2}(t) + ... + \beta_k x_{jk}(t)) \quad (3)$$

Hazard ratio between the two observations at time $'t'$ is defined as

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t|1)}{h_0(t|0)} = \frac{h_0(t)\exp\left[\sum_{i=1}^{n}\beta_i x_i\right]}{h_0(t)\exp\left[\sum_{j=1}^{n}\beta_j x_j\right]} = \exp(\sum_{i=1}^{n}\beta_i\left(x_i - x_j\right)) \quad (4)$$

where $'t'$ independent of the hazard ratio of the covariates at any given time and h_0 is hereby removed; but not on baseline hazard. Cox(1972) reported partially likelihood function and maximum likelihood function to estimate unspecified baseline hazard. Otherwise the following distributions exponential, Weibull fully parametric CPH is obtained since baseline hazard is known.

Maximum partially likelihood function to estimate β

$$L(\beta) = \prod_{v\in\Delta}\frac{\exp\left[\sum_{i=1}^{n}\beta_i x_i\right]}{\sum_{r\in R(t_v)}\exp\left[\sum_{j=1}^{m}\beta_j x_j\right]} \quad (5)$$

The log likelihood function of equation

$$\pi(\beta) = \sum_{v\in\Delta}\sum_{i=1}^{n}\beta_i x_i - \sum_{v\in\Delta} In\left(\sum_{r\in R(t_v)}\exp\left(\sum_{i=1}^{n}\beta_i x_{ri}\right)\right) \quad (6)$$

First derivative (Score function)

$$\pi'(\beta) = \sum_{v} x_v \sum_{i\in\Delta}\frac{\sum_{r\in R(t_v)} x_r \exp\left[\sum_{i=1}^{n}\beta_i x_i\right]}{\sum_{r\in R(t_v)}\exp\left[\sum_{j=1}^{m}\beta_j x_j\right]} \quad (7)$$

Second derivative with $c, g = 1, 2, ..., n$

$$\pi''(\beta) = \pi''_{(cg)}(\beta) = \sum_{i\in\Delta}\frac{\sum_{r\in R(t_v)} x_{ir}x_{cp}\exp\left[\sum_{i=1}^{n}\beta_i x_{ir}\right]}{\sum_{r\in R(t_v)}\exp\left[\sum_{i=1}^{n}\beta_i x_{ir}\right]} - \sum_{i\in\Delta}\left[\frac{\sum_{r\in R(t_v)} x_{gr}\exp\left[\sum_{i=1}^{n}\beta_i x_{ir}\right]}{\sum_{r\in R(t_v)}\exp\left[\sum_{i=1}^{n}\beta_i x_{ir}\right]}\right]\left[\frac{\sum_{r\in R(t_v)} x_r\exp\left[\sum_{i=1}^{n}\beta_i x_{ir}\right]}{\sum_{r\in R(t_v)}\exp\left[\sum_{i=1}^{n}\beta_r x_{ir}\right]}\right] \quad (8)$$

Then

Estimate is given as follows

$$\hat{h}(\beta) = \frac{\sum_{i=1}^{n} x_i t\exp\left(x_i t(\beta)\right)}{\sum_{i=1}^{n} x_i t\exp\left(x_i t(\beta)\right)} = \sum_{i=1}^{n} x_i w_i \quad (9)$$

$\hat{h}(\beta)$ is called weighted average of the covariates (x) of among subject at risk at time (t).

$$w_i = \frac{\exp\left(x_i t(\beta)\right)}{\sum_{i=1}^{n} x_i t\exp\left(x_i t(\beta)\right)} \quad (10)$$

$$V_x(\beta) = \frac{\sum_{i=1}^{n}\left(x_i t\right)^2\exp\left(x_i t(\beta)\right)}{\sum_{i=1}^{n} x_i t\exp\left(x_i t(\beta)\right)} - \left(\frac{\sum_{i=1}^{n} x_i t\exp\left(x_i t(\beta)\right)}{\sum_{i=1}^{n} x_i t\exp\left(x_i t(\beta)\right)}\right)^2 \quad (11)$$

$$V_x(\beta) = \frac{\sum_{i=1}^{n}(x_i t)^2 \exp(x_i t(\beta))}{\sum_{i=1}^{n} x_i t \exp(x_i t(\beta))} - \left(\hat{h}(\beta)\right)^2 \quad (12)$$

$$V_x(\beta) = \frac{\sum_{i=1}^{n}(x_i t)^2 \exp(x_i t(\beta))}{\sum_{i=1}^{n} x_i t \exp(x_i t(\beta))} - \left(\frac{\sum_{i=1}^{n} x_i t \exp(x_i t(\beta))}{\sum_{i=1}^{n} x_i t \exp(x_i t(\beta))}\right)^2 \quad (13)$$

$$V_x(\beta) = \sum_{i=1}^{n} \frac{\sum_{i=1}^{n}\left(x_i t - \hat{h}(\beta)\right)^2 \exp(x_i t(\beta))}{\sum_{i=1}^{n} x_i t \exp(x_i t(\beta))} = \sum_{i=1}^{n}\left(x_i t - \hat{h}(\beta)\right)^2 w_i \quad (14)$$

$V_x(\beta)$ is the weighted variance of the covariate (x) among subject at risk at time (t).

Simulation Study

Simulation experiments were conducted to investigate the performance of CPH model with correlated unordered categorical data. Two Unordered categorical data (x_1 and x_2) were generated using binomial distribution. When the true regression coefficients are $\beta_1 = \beta_2 = 1$, the datasets for x_1 and x_2 over time to event, correlated at ρ =0.2,0.6 and 0.9 for sample size ranges from 10-500. When $T_i = t$ and $T_i > t$, $\delta_i = 1$ and 0 are specified and unspecified entry study period respectively study period denoted by t. The right censoring($T_i \geq 0$) and baseline hazard($\lambda_0(t)$) were generated from uniform and exponential distribution respectively. In this form, data were generated from Survival time, $T = H_0^{-1}\left(\frac{-\log U}{\exp(x'\beta)}\right)$, where cumulative hazard function, $H_o(t) = \lambda(t)$ follows exponential distribution with $\lambda(t) = 0.5z_1 + 0.5z_2$, where z_1, z_2 follow log normal distribution and its inverse is given as $H_o^{-1} = \lambda(t)^{-1}$.

Table 1. Estimates of in Time Independent Cox Proportional Hazard Model with Correlated Unordered categorical Data

Discussion of Results

Table 1 shows the estimates, Hazard ratios and p-values of the coefficients at 5% level of significance when the model parameters, $\beta_1 = \beta_2 = 1$ with varied collinearity values(ρ) for

Collinear(ρ)		True value($\beta_1 = 1$)				True value($\beta_2 = 1$)			
		10	30	100	500	10	30	100	500
$\rho = 0.2$	Coefficients	2.3210	0.4563	0.7481	0.9558	1.960	0.7291	1.2318	1.04225
	Se	1.8190	0.3866	0.2167	0.0973	0.3866	1.111	0.2269	0.09979
	Hazard	1.2050	1.578	2.1131	2.6008	7.099	2.0732	3.4275	2.83559
	Ratios	0.9990	0.2380	0.0006	0.000*	0.0776	0.0664	0.0000*	0.0000*
	P-values			*					
$\rho = 0.6$	Coefficients	2.2970	0.5073	0.6644	0.9079	2.149	0.8937	1.3084	1.1464
	Se	1.6720	0.3857	0.2177	0.1039	1.136	0.4030	0.2389	0.1087
	Hazard	9.4820	1.6609	1.9433	2.4791	8.576	2.4443	3.7003	3.1468
	Ratios	0.9989	0.1884	0.0022	0.000*	0.0585	0.0266*	0.0000*	0.0000*
	P-values			*					
$\rho = 0.9$	Coefficients	2.3560	1.3115	0.3633	1.0965	2.345	-0.3132	1.5716	0.8569
	Se	2.4140	0.6826	0.3087	0.2071	1.192	0.6487	0.3476	0.2094
	Hazard	1.702	3.7116	1.4381	2.9935	1.043	0.7311	4.8142	2.3558
	Ratios	0.9992	0.0547	0.239	0.000*	0.049*	0.6292	0.0000*	0.0000*
	P-values								

sample size(N) ranges from 10-500 were obtained through simulation.

The estimates obtained when $\beta_1 = 1$ are 2.3210(1.8190), 2.2970(1.6720), 2.3560(2.4140) and when $\beta_2 = 1$ are 1.960(0.3866), 2.149(1.136), 2.345(1.192) both at N=10 for ρ, 0.2,0.6 and 0.9 respectively. The hazard ratio of $\beta_1(\beta_2)$ when N=10 at ρ =0.2,0.6 and 0.9 are 1.2050(7.099), 9.4830(8.576) and 1.702(1.043) with P-values of 0.9990(0.0776),0.9989(0.00585) and 0.9992(0.049) at 5% level of significance respectively. Also, when N=100 and 500 for both β_1(0.2167, 0.0973) and β_2 (0.2269, 0.09979) respectively, the standard errors of β_1 and β_2 at ρ =0.2 are smaller compared to that of ρ =0.6 and 0.9. Findings reveal that as collinearity becomes large, the coefficients are not close to the true values for all considered collinearity values. There is statistically significant association between x_2 and time-to- event at ρ =0.9 with β_2 = 2.345(P- value = 0.049) in the model. From the Table1, the standard errors become large when ρ =0.2 and 0.9 as against 0.6. Thus, for large sample sizes at low collinearity value, ρ =0.2, the correlated unordered categorical covariate(CC) have statistically significant association with time-to-event with smaller standard errors than that of small sample sizes. The hazard ratios of regression coefficients of x_1 and x_2 are greater than one for all Ns considered except when N=30 for x_2. The effects of collinearity for unordered CC in Cox model are risky for interpretation and hazard assumption. This implies that there is a strong evidence of non proportional hazard assumptions for the coefficients of x_1 and x_2 for where the values appeared with asterisk(*) in the Table 1.

Table 2. Model Diagnostic for Cox Proportional Hazard at 5% level of significance

	10	30	100	500
$\rho = 0.2$	0.0030	0.1000	0.0000	0.0000
$\rho = 0.6$	0.0020	0.0400	0.0000	0.0000
$\rho = 0.9$	0.0020	0.0400	0.0000	0.0000

Table 2 shows the P values at 5% level of significance of Likelihood ratio test for testing Cox model fitness under different levels of collinearity(0.2,0.6 ,0.9) as sample size(N) increases. When $\rho = 0.2$, the obtained P values are 0.0030, 0.1000 for N of 10 and 30 respectively; equal p value=0.0000 was obtained for sample sizes of 100 and 500. This implies that the model fitted is statistically significant when all sample sizes considered except when N= 30. It is also evident in Table 2 that the model is statistically significant at the considered levels of collinearity($\rho = 0.6, 0.9$) for all sample sizes.

Fig1, 2 and 3 show the graphical representation of both significant and insignificant effects of the collinearity on the ordered CC over time. The graphs are facilitated by smoothing spline shown on the graphs by horizontal lines while the broken lines represent the ± 2 standard errors around the fit. The departures from the horizontal lines represent the non proportional hazard. Fig1a, 2a, and 3a show the significant effect of collinearity on covariate(x_1) is constant over time. Hence, there is no violation to proportional hazard assumptions when there exists collinearity between unoredered CC for small sample size. However, there is a very strong evidence of departures from the horizontal lines in Fig 1, 2 and 3 (b,c,d,e,f,g and h) for N=30,100,500 for all considered values of collinearity. These graphs indicate the non-proportional hazard assumptions and that effects of collinearity on the covariate(x_1 and x_2) are not constant over time. With insufficient reasoning behind the collinearity between unordered categorical data with time independency of CPH model, this paper contributes to literature by examining this assumption violation and its negative effects against hypothesis testing of zero regression coefficient for continuous covariate data that are primarily grouped(Finkelstein, 1986). This study is in contrary to effects of collinearity on the estimates of time dependent coefficient in CPH model by Babalola and Yahya(2019).

Conclusion

This paper estimates CPH model with correlated unordered CC. The effects of correlated Binary Independent Variables(BIVs) has not gained much needed attention theoretically and empirically in Cox model. Findings reveal that the coefficients with reduced standard errors obtained are close to the true values for large sample size at different level of collinearity. This study has shown the effects of collinearity between two unordered CC on the performance of cox model for different sample size. In addition, there is significant effects of collinearity on unordered CC which is a strong indication of non-proportional hazard assumptions in the model. Consequentially, the effects of collinearity between unordered CCs violate the PH assumption which consequentially leads to problem of overestimation and underestimation.

References

Cox, D. R.(1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B, 34, 187-220.

Babalola B. T. and Yahya W. B. (2019).Effects of Collinearity on Cox Proportional Hazard Model with Time Dependent Coefficients: A Simulation Study J Biostat Epidemiol ;5(1): 172-182

Azimmatul Ihwah(2015).The Use of Cox Regression Model to Analyze the Factors that Influence Consumer Purchase Decision on a Product. The 2014 International Conference on Agro-industry (ICoA) : Competitive and sustainable Agroindustry for Human Welfare. ESELVIER, Agriculture and Agricultural Science Procedia 3 (2015) 78 – 83

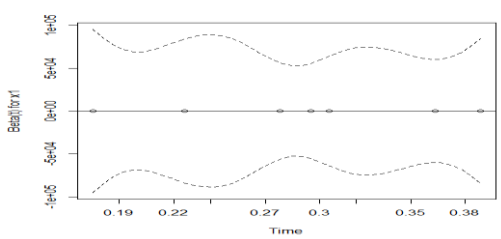
Triastuti W, Sri H. Kartiko, Danardono(2018).The Cox proportional Hazard model on

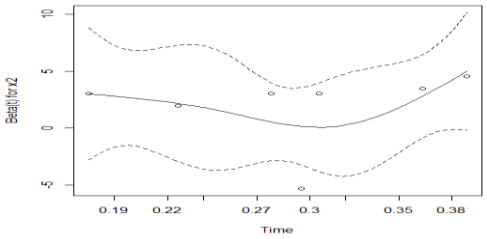
duration of birth process. IOP Conf. Series: Journal of Physics: Conf. Series 1025 - 012121, DOI :10.1088/1742-6596/1025/1/012121

Xue X., Kim M. Y, Shore R. E. (2007).Cox regression analysis in presence of collinearity: an application to assessment of health risks associated with occupational radiation exposure. Lifetime Data Anal, 13:333–350 DOI 10.1007/s10985-007-9045-1

Finkelstein Dianne M. (1986).A Proportional Hazards Model for Interval-Censored Failure Time Biometrics, Vol. 42, No. 4, pp. 845-854

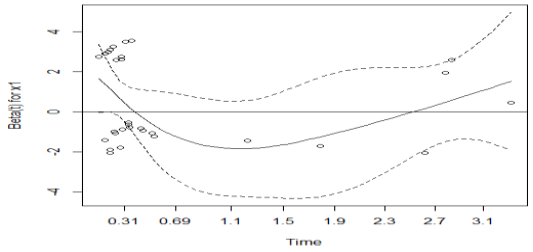
Ledwon , Andreas V. and Jäger , Clemens C. (2020).Cox Proportional Hazards Regression Analysis to assess Default Risk of German-listed Companies with Industry Grouping. ACRN Journal of Finance and Risk Perspectives journal homepage: http://www.acrn-journals.eu/

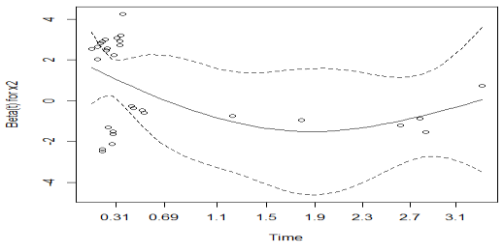




1a.Effect of Collinearity on β_1 1b. Effect of Collinearity on β_2

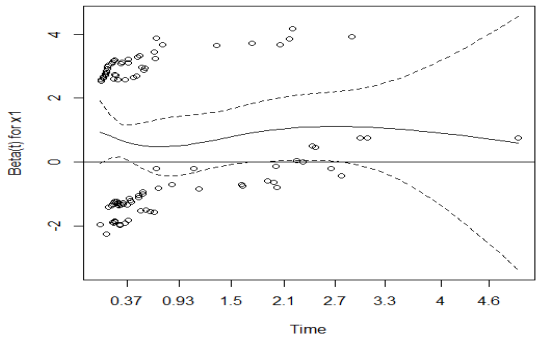
Fig1. Graphical representations of the effects of collinearity($\rho = 0.2$) on the correlated unordered categorical data for sample sizes(N=10)

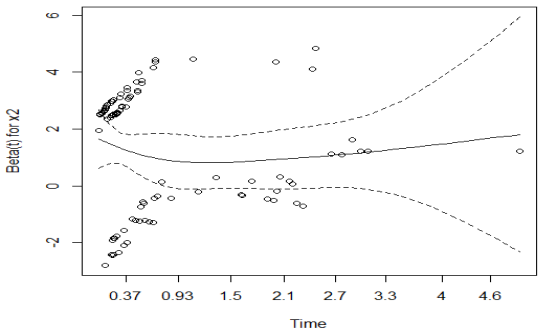




1c.Effect of Collinearity on β_1 1d. Effect of Collinearity on β_2

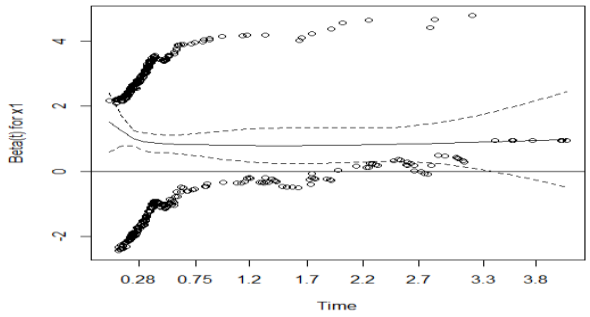
Fig1. Graphical representations of the effects of collinearity($\rho = 0.2$) on the correlated unordered categorical data for sample sizes(N=30)

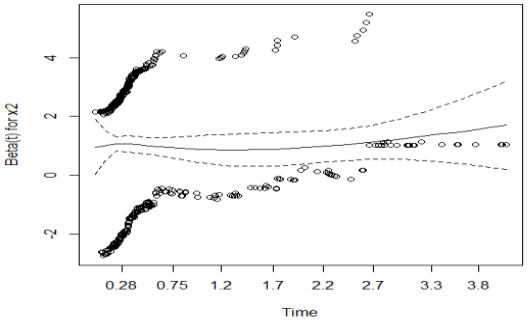




1e.Effect of Collinearity on β_1 1f. Effect of Collinearity on β_2

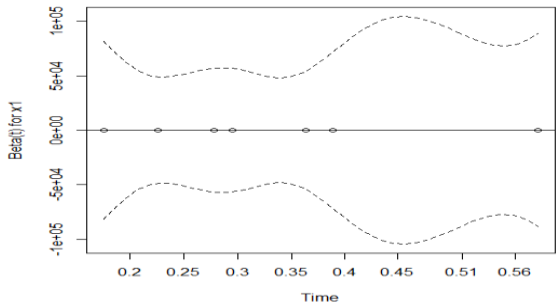
Fig1. Graphical representations of the effects of collinearity($\rho = 0.2$) on the correlated unordered categorical data for sample sizes(N=100)

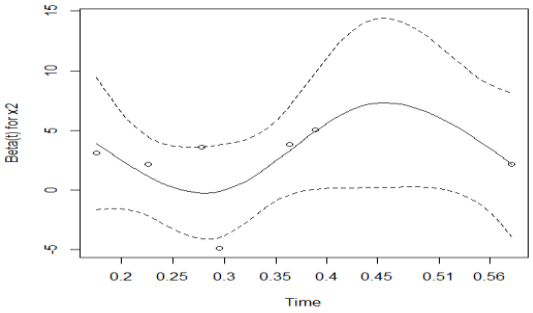




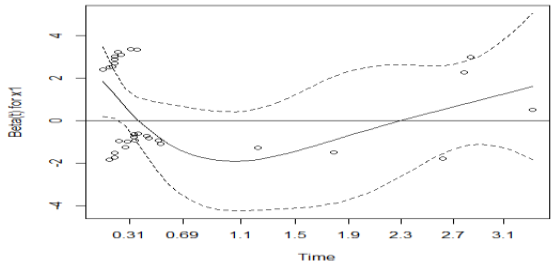
1g.Effect of Collinearity on β_1 1h. Effect of Collinearity on β_2

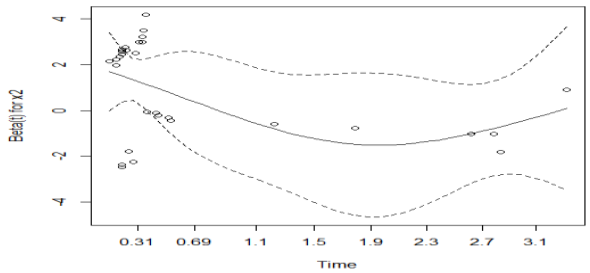
Fig1. Graphical representations of the effects of collinearity($\rho = 0.2$) on the correlated unordered categorical data for sample sizes(N=500)





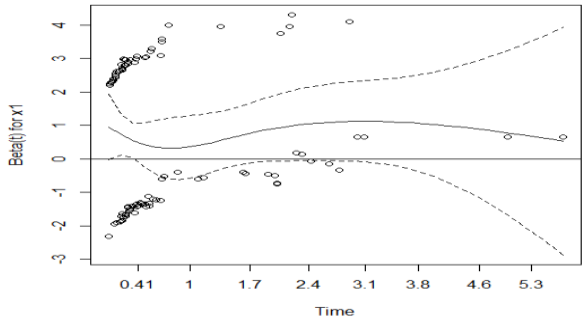
2a.Effect of Collinearity on β_1 2b. Effect of Collinearity on β_2 **Fig2.Graphical representations of the effects of collinearity($\rho = 0.6$) on the correlated unordered categorical data for sample sizes(N=500)**

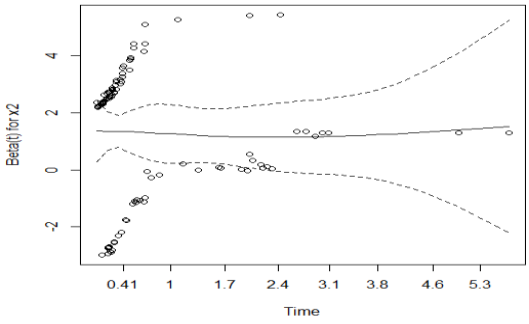




2c.Effect of Collinearity on β_1 2d. Effect of Collinearity on β_2

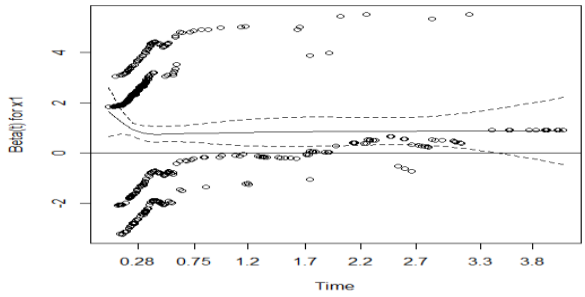
Fig2.Graphical representations of the effects of collinearity($\rho = 0.6$) on the correlated unordered categorical data for sample sizes(N=30)

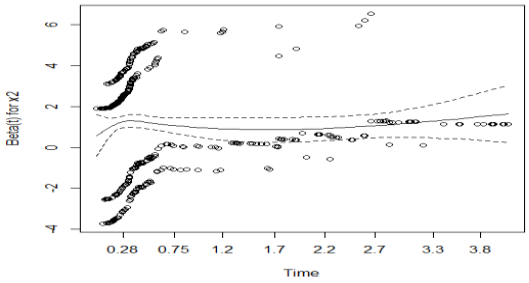




2e.Effect of Collinearity on β_1 2f. Effect of Collinearity on β_2

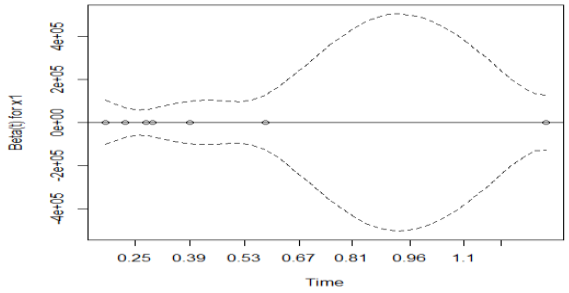
Fig2.Graphical representations of the effects of collinearity($\rho = 0.6$) on the correlated unordered categorical data for sample sizes(N=100)

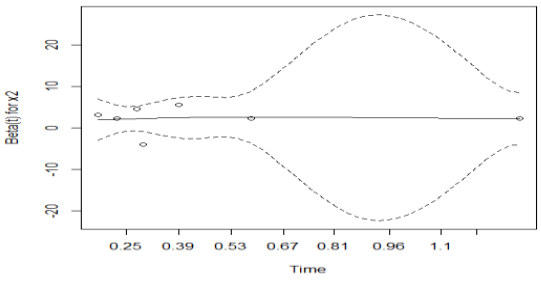




2g.Effect of Collinearity on β_1 2h. Effect of Collinearity on β_2

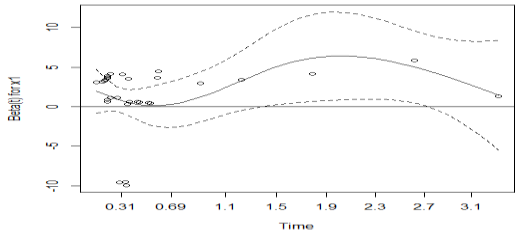
Fig2.Graphical representations of the effects of collinearity($\rho = 0.6$) on the correlated unordered categorical data for sample sizes(N=500)

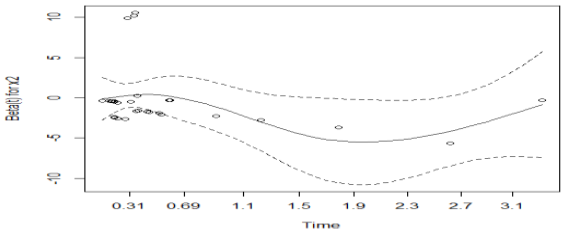




3a.Effect of Collinearity on β_1 a 3b. Effect of Collinearity on β_2

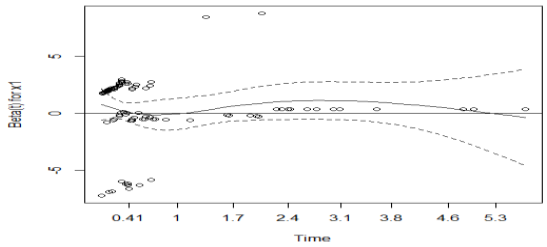
Fig3. Graphical representations of the effects of collinearity($\rho = 0.9$) on the correlated unordered categorical data for sample sizes(N= 10)



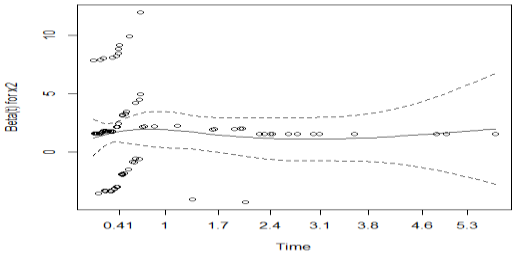


3c.Effect of Collinearity on β_1 3d. Effect of Collinearity on β_2

Fig3. Graphical representations of the effects of collinearity($\rho = 0.9$) on the correlated unordered categorical data for sample sizes(N= 30)

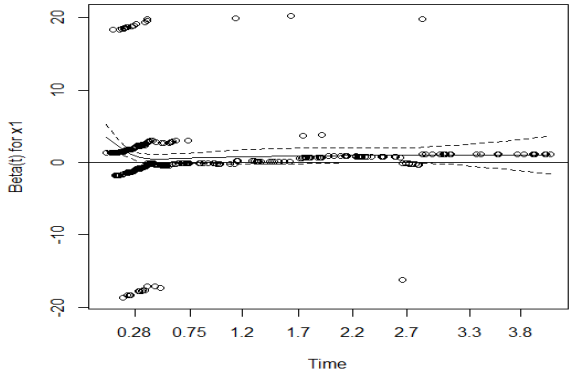


3e.Effect of Collinearity on β_1

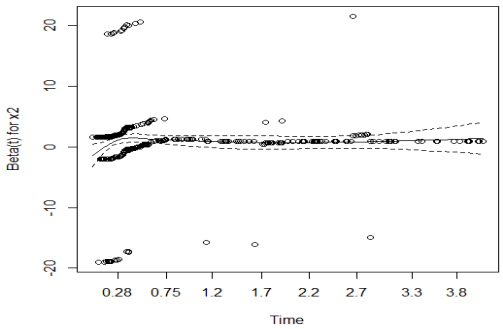


3f. Effect of Collinearity on β_2

Fig3. Graphical representations of the effects of collinearity($\rho = 0.9$) on the correlated unordered categorical data for sample sizes(N= 100)



3g.Effect of Collinearity on β_1



3h. Effect of Collinearity on β_2

Fig3. Graphical representations of the effects of collinearity($\rho = 0.9$) on the correlated unordered categorical data for sample sizes(N= 500)