Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2022, 2022:89

https://doi.org/10.28919/cmbn/7623

ISSN: 2052-2541

EXAMINING FACTORS AFFECTING DELAYED COMPLETION OF ADJUVANT CHEMO FOR PATIENTS WITH BREAST CANCER: DEVELOPMENT OF RIDGE LOGISTIC PANEL ESTIMATORS

AMERA M. EL-MASRY[1], AHMED H. YOUSSEF[2], MOHAMED R. ABONAZEL[2,*]

[1]Department of Mathematics and Statistics, Faculty of Management Technology and Information Systems, Port Said University, Port Said, Egypt

[2]Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research (FGSSR), Cairo University, Giza, Egypt

Abstract: The problem of multicollinearity among predictor (independent) variables is a frequent issue in logistic panel data analysis. The model parameters are estimated via the conditional maximum likelihood and unconditional maximum likelihood estimators. In this context, this paper proposes a ridge regression estimation via shrinkage methods to analyze such data. Furthermore, in view of obtaining more efficient estimators, we propose ridge estimators using different shrinkage parameters for the fixed effects logistic panel data model. An application is also presented to assess the performance of the proposed ridge estimators. The most significant factors that affect delayed completion of adjuvant chemotherapy in patients with breast cancer plus their existing outcomes in order to shed light on the link between chemotherapy duration and its outcomes according to breast cancer are illustrated in the study. The study results show that the conditional fixed effects logit estimator is more efficient and better than the unconditional pooling

*Corresponding author

E-mail address: mabonazel@cu.edu.eg

Received July 23, 2022

and unconditional fixed effects logit estimators. Moreover, we find that there are very influential factors that affected delayed completion of adjuvant chemotherapy such as Body Surface Area (BSA), Hemoglobin (HGB), Alanine Transaminase (ALT) and Creatinine (SRCR).

Keywords: panel data; logistic regression; fixed effects; conditional maximum likelihood; multicollineairty; ridge regression.

2010 AMS Subject Classification: 62J07, 62J12, 62P10.

1. INTRODUCTION

In general, the panel data models are better suited to study the dynamics of change and that captures the statistical relationship between the dependent and independent variables, see, e.g., [1, 2, 3, 4, 5, 6, 7, 8]. In logistic panel data setup, the response (dependent) variable is a binary choice variable taking values 1 or 0 for success or failure respectively. Repeated measures of some variables of interest are collected over a specified time for different individuals [9]. These types of data are repeatedly found in medical research where the responses are influenced by different time-dependent and time-constant factors. It is quite natural that the repeated measures shall exhibit some problems, as the data in the model are dependent, namely multicollineairty problem, a problem that arises in situations when the covariate (independent) variables are high inter-correlated [10, 11, 12]. Then it becomes difficult to disentangle the separate effects of each of the covariates variables on the response variable. Thus, the estimated parameters may be insignificant on the statistical basis and/or have different signs without any expectation. Thus, conducting a meaningful statistical inference would be difficult for the researcher. The ridge regression estimator can improve the estimation of β by adding a small constant to the diagonal of the matrix, see [10, 11, 12, 13, 14, 15, 16].

Schaefer et al. [12] is followed to elaborate ridge regression theory in logistic regression related to logistic panel regression. Drawing on the similarities between the logistic regressions and logistic panel data model, this paper proposed some new methods of estimating the shrinkage parameter to be used in RLP in order to combat multicollinearity in binary logistic panel data regression

model. RLP with new estimators are expected to perform better than conditional maximum likelihood when the explanatory variables are correlated. Moreover, we give a matrix mean squared error (MSE) comparison between the estimators and conduct an application study based on breast cancer data to evaluate the performances of the estimators using the MSE.

In this regard, the breast cancer is the most well-known type of cancer among women around the world, as it represents 16% of all cancers that affect this category. Every year there are about 1.38 million new cases of Breast Cancer and 458,000 deaths from Breast Cancer (according to estimates by the Globocan website 2008 of the International Agency for Research on Cancer). Although some believe that this cancer is a disease of the developed world, most 69% deaths occur in developing countries, as well. In recent years, cancer rates have been shown to be raising steadily in low-and middle income countries (according to the WHO global burden of disease report). Chemotherapy presented for Breast Cancer uses anti-cancer drugs that may be used intravenously (injected into your vein) or orally. The needed drugs travel through the bloodstream to target cancer cells in many parts of the body. Chemotherapy drugs prescribed to treat Breast Cancer can be presented before surgery (neoadjuvant) or after surgery (adjuvant). After surgery (adjuvant chemotherapy), Adjuvant chemo might be given to try to kill any cancer cells that might have been left behind or have spread but cannot be seen, even on imaging tests. If these cells could grow, they could form new tumors in other places in the body. Adjuvant chemo can lower the risk of Breast Cancer not to appear again. Before surgery (neoadjuvant chemotherapy), Neoadjuvant chemo could work out to minimize the tumor to be removed with the least extensive surgery.

For this reason, neoadjuvant chemo is often functioned to treat cancers to be removed by surgery when diagnosed for the first time to be classified as locally advanced cancers. For certain types of Breast Cancer, there are tumor cells, if found at the time of surgery, that are called residual disease. Many patients may be offered more chemotherapy after surgery to reduce the chances of the cancer not to come back (recurrence). For types of chemicals used, there are two types: Adriamycin Cyclophosphamide (A/C) and Taxol. In this study, surgical adjuvant Chemotherapy will be highlighted after mastectomy for Breast Cancer to determine the most important factors affecting

delayed completion of adjuvant chemotherapy among patients. The following Reviews of Literature for application are presented as follows.

Nissen-Meyer et al. [17] proposed the importance of chemotherapy after the surgery as divided into three patient sections. First: One single six-day course with cyclophosphamide (total dose 30 mg/kg) was given immediately after mastectomy to 507 breast cancer patients. Second: 519 randomized controls received no adjuvant chemotherapy. Third: other breast cancer patients received chemotherapy course of three weeks after mastectomy. Therefore, after such chemotherapy, the control group then has 234 recurrences and 196 deaths, and the treatment group 175 recurrences and 146 deaths. The differences of fifty deaths in favor of the treatment group are significant. In this regard, the differences in recurrence rates increased step by step to reach 10.71% four years after mastectomy, and to be fixed for another six years. The rates of differences in death increased for six years after mastectomy to be 10.48% after 10 years. This pattern has the mechanism of a lacking delay in onset of clinical recurrences with an absolute reduction of recurrence rates because of tumoricidal chemotherapy. The same chemotherapy course given three weeks after mastectomy seemed without effect.

Bleiker et al. [18] used conditional logistic regression analysis to identify variables that could best explain group membership, i.e., belonging to the case (Breast Cancer) or the control (without disease) group, to determine the factors that affect how far Breast Cancer develops. They used longitudinal study design from 1989 through 1990. For that purpose, a personality questionnaire was sent to all female residents of the Dutch city of Nijmegen who were forty-three years of age or older to investigate these significant personality factors in addition to somatic risk factors, he may be considered with the development of primary Breast Cancer. Results: For personality to be used for statistical analyses purpose, three variables were found to be statistically significant to be associated with an increased risk of Breast Cancer: 1) The existence of a first-degree family member with Breast Cancer, 2) nulliparity and 3) a relatively high score on the personality scale of anti-emotionality.

Chavez-MacGregor et al. [19] used logistic regression and Cox proportional hazard models in

order to highlight the determinants in delayed chemotherapy initiation of adjuvant chemotherapy in Patients with Breast Cancer and to identify the link between Time To Chemotherapy (TTC) and its existing outcomes related to Breast Cancer subtype. With the help of data from the California Cancer Registry, we could study a number of 24843 patients diagnosed with Breast Cancer in the period between 2005 and 2010 to be treated with adjuvant chemotherapy. Results: Those factors related to delays in TTC contained low socioeconomic status, nonprivate insurance and Hispanic ethnicity or Non-Hispanic black race. When compared with patients who receive chemotherapy within thirty-one days from surgery, there was no evidence of existing adverse outcomes in those with TTC of thirty-one to sixty or sixty to ninety days. Patients treated ninety-one or more days from surgery experienced a worse overall survival.

Bray et al. [20] delivered a status report on the global burden of cancer worldwide utilizing the GLOBOCAN 2018 estimates of cancer incidence and mortality set by the International Agency for Research on Cancer with a concentration on geographic variability across twenty regions worldwide. In this regard, there will be an estimated 18.1 million new cancer cases (17.0 million excluding nonmelanoma Skin Cancer) and 9.6 million cancer deaths (9.5 million excluding nonmelanoma Skin Cancer) in 2018. Lung Cancer turns out to be the most commonly diagnosed cancer in both sexes combined of 11.6% of the total cases, and the leading cause of cancer deaths (18.4% of the total cancer deaths) was closely followed by female Breast Cancer (11.6%), Prostate Cancer (7.1%) in addition to Colorectal Cancer (6.1%), for incidence, Colorectal Cancer (9.2%), Stomach Cancer (8.2%), and Liver Cancer (8.2%) for mortality. The international initiative for cancer registry development is an international partnership that grants better estimation in addition to the collection and the use of local data to prioritize and evaluate national cancer control efforts. This paper is organized as follows: Section 2 is to focus on methodology and the proposed ridge estimator. The estimation methods of the ridge parameter are presented in Section 3. Details of the empirical study are given in Section 4. Finally, we provide a brief conclusion in section 5.

2. METHODOLOGY

2.1. Fixed Effects Logit Panel Data (FELPD) Model

In many economic studies, the response variable is categorical indicating a success or a failure of an event. Such a dependent variable is normally represented by a binary choice variable $y_{it} = 1$ if the event happens and 0 if it does not happen of individual i at time t. Consider the non-linear binary response model as:

$$pr(y_{it} = 1;\ \alpha_i, \beta) = G(\alpha_i + x_{it}\beta) = \mu_{it} \tag{1}$$

$$pr(y_{it} = 0;\ \alpha_i, \beta) = 1 - \mu_{it} \tag{2}$$

Where $G(\alpha_i + x_{it}\beta)$ is a nonlinear function taking on values strictly between zero and one: $0 < \mu_{it} < 1$,various non-linear functions for G have been suggested in the literature by far the most common ones to be the logistic distribution, yielding the logit model. The logit model takes the following form:

$$\mu_{it} = \frac{exp(\alpha_i + x_{it}\beta)}{1 + exp(\alpha_i + x_{it}\beta)};\ i = 1, .., N;\ t = 1, \dots, T \tag{3}$$

Where y_{it} is the response, and in the case of a logistic panel model, a binary response variable is an indicator for individual *i* at time *t*. Such that $y_{it} = 1$ if an event occurs and $y_{it} = 0$ if it does not occur. This is the CDF for a logistic variable, where x_{it} is row vector of the observed covariates variables, β is a vector of parameters, α_i is an unobserved time invariant individual effect [21].

For estimation, logistic panel data method has been used throughout applying two models were pooled logistic regression model is estimated by maximum likelihood estimator. When the panel data structure and the response variable are binary with this method, the panel structure of the data is ignored.

The FELPD models are estimated by two estimation methods by the unconditional maximum likelihood (UCML) and conditional maximum likelihood (CML) estimators. The Hausman specification test compares UCML and CML estimators, where the null hypothesis is the UCML and CML estimators are consistent. In this regard CML is inefficient, whereas the alternate hypothesis UCML is inconsistent and inefficient with CML being consistent and efficient.

2.2. Model Assumptions

Assumption (1): The probability of observing $Y_{it} = 1$ is $G(\alpha_i + x_{it}\beta)$ while the probability of observing $Y_{it} = 0$ is $1 - G(\alpha_i + x_{it}\beta)$.

Assumption (2): The true conditional probabilities are logistic function of the explanatory variables. $pr(y_{it} = 1)$ depends on x_{it} through the logistic function.

Assumption (3): The covariates variables are not linear combinations of each other.

Assumption (4): The covariates variables are measured without error.

Assumption (5): No important variables are omitted, and no extraneous variables are included.

Assumption (6): Conditional on x_{it}, y_{it} is an independent Bernoulli random variable with probability given by (3).

Assumption (7): $y_{i1}, \dots, y_{it}$ are independent conditionals on (α_i, x_{it}).

Assumption (8): The conditional probability that y equals one is equal to conditional expected value of y_{it}, i.e., $pr(y_{it} = 1; \alpha_i, x_{it}) = E(y_{it}; \alpha_i, x_{it})$.

2.3. Conditional Fixed Effects Logit Estimator

The CML estimator of the FELPD model is usually called the "conditional fixed effects logit" estimator, as we must emphasize that the fixed effects logit estimator does not arise by treating the α_i as parameters to be estimated along with β, see [22].

The conditional likelihood approach can be applied directly to the FELPD model, since $\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for α_i. This conditional likelihood function does not depend upon α_i. The conditional likelihood function is in the form of a binary logit likelihood function in which the two response are (0, 1) and (1, 0) with covariates variables. The CML estimate of β can be obtained simply from a standard maximum likelihood binary logit estimation. The general presentation of this model is quite complex, but the intuition of it can be perceived using the special case where $T = 2$. Consider first the case of $T = 2$ if $y_{i1} + y_{i2} = 0$ or 2 then y_{i1} and y_{i2} are both determined given their sum. So the only case of interest is $y_{i1} + y_{i2} = 1$.

Then the two possibilities are $w_i = 1\ if\ (y_{i1} = 0, y_{i2} = 1) = (0,1)$ and $w_i = 0\ if\ (y_{i1} =$

$1, y_{i2} = 0) = (1,0)$. The conditional density is:

$$pr(w_i = 1; y_{i1} + y_{i2} = 1) = \frac{e^{\beta'(x_{i2}-x_{i1})}}{1+e^{\beta'(x_{i2}-x_{i1})}} = G[\beta'(x_{i2} - x_{i1})] \quad (4)$$

Which does not depend on α_i.

The conditional log-likelihood function is:

$$L^c = \sum_{i\in I_1}\{w_i \log\{G[\beta'(x_{i2} - x_{i1})]\} + (1 - w_i) \log\{G[-\beta'(x_{i2} - x_{i1})]\}\} \quad (5)$$

Where $I_1 = \{1; y_{i1} + y_{i2} = 1\}$. The conditional maximization likelihood obtains consistent estimates of β [23, 24]. One should take the derivative of it with respect to β and set the resulting equations called the likelihood equations to zero. Since the resulting equation is nonlinear in parameters, so some special methods should be used in order to obtain the solution. The iteratively re-weighted least squares (IRLS) method can be applied to get the solution. Then the maximum likelihood estimator CML of β can be obtained by using IRLS algorithm as follows:

$$\hat{\beta}_{CML} = \left(X'\widehat{\Psi}_{LP}X\right)^{-1}\left(X'\widehat{\Psi}_{LP}\hat{Z}\right) \quad (6)$$

Where X is $(NT \times p)$ of observed explanatory variables, $\widehat{\Psi}_{LP}$ is the weighting matrix for the logistic panel data model; $\widehat{\Psi}_{LP} = \text{diag}\left(\hat{\mu}_{\text{it}}\left(1 - \hat{\mu}_{\text{it}}\right)\right)$, $\hat{Z} = (\hat{z}_{11}, \dots, \hat{z}_{NT})'$; $\hat{z}_{it} = \log\left(\hat{\mu}_{\text{it}}\right) + \frac{\text{y}_{\text{it}} - \hat{\mu}_{\text{it}}}{\hat{\mu}_{\text{it}}\left(1-\hat{\mu}_{\text{it}}\right)}$.

The hats in the equation show the iterative process. Hence, use this to update the estimate of β until convergence. Then the trace mean squared error (TMSE) of the CML is

$$TMSE\left(\hat{\beta}_{CML}\right) = E\left(\hat{\beta}_{CML} - \beta\right)'\left(\hat{\beta}_{CML} - \beta\right) = tr\left(X'\widehat{\Psi}_{LP}X\right)^{-1} = \sum_{j=1}^{p}\left(\frac{1}{\lambda_j}\right) \quad (7)$$

Where λ_j is the j^{th} eigenvalue of the $\left(X'\widehat{\Psi}_{LP}X\right)$ matrix. One of the disadvantages of using CML is that its MSE becomes inflated when the covariate variables are highly inter-correlated which is called the multicollinearity problem when the columns of the matrix X are dependent. As a consequence of collinearity between the independent variables, some of the eigenvalues of the matrix $\left(X'\widehat{\Psi}_{LP}X\right)$ become very close to zero. Thus, the variance of CML becomes so large that the estimations become unstable. As a solution to this problem, see [10, 11] for more details on ridge regression for the ordinary least square situation, Schaefer et al. [12] proposed the following logistic version of the ridge estimator (LRE).

2.4. Proposed Estimator: Fixed Effect Ridge Regression

Schaefer et al. [12] is followed to elaborate ridge regression theory in logistic regression in relation to logistic panel regression. Out of the similarities between the logistic regressions and logistic panel data model, ridge parameter estimator is suggested in this regard. The ridge logistic panel estimator is the outcome as a limited maximum likelihood estimator.

Consider the maximization of the log-likelihood function with a penalty on the norm of β:

$$L^{pen}(y_{it}, x_{it}; \beta, k) = L(y_{it}, x_{it}; \beta) - \frac{1}{2}k\beta'\beta \tag{8}$$

Where $L(y_{it}, x_{it}; \beta)$ is the log likelihood function, $k > 0$ is penalty parameter. The ridge parameter k controls the amount of shrinkage of the norm of β. When $k = 0$, the solution of this formula will be the well-known CML. A large number of covariate variables and/or much correlation between the various covariate variables give rise to unstable parameter estimates. Shrinking the β towards 0 and allowing a little bias will stabilize the system to provide estimates with smaller variance. Therefore, for a good choice of k, the estimate $\hat{\beta}(k)$ is expected to be on an average closer to the real value of β than the unrestricted CML, i.e. MSE ($\hat{\beta}(k)$) < MSE ($\hat{\beta}_{CML}$). Using the Newton-Raphson algorithm for solving the (penalized) estimating equation, we get the ridge conditional maximum likelihood (RCML) estimator:

$$\hat{\beta}(k) = \hat{\beta}_{RCML} = \left(X'\widehat{\Psi}_{LP}X + kI_p\right)^{-1}\left(X'\widehat{\Psi}_{LP}\hat{Z}\right) \tag{9}$$

2.5. Properties of the Proposed Estimator

The efficiency of an estimator is properly quantified by the MSE. Generally speaking, for any estimator of a parameter θ:

$$MSE\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + \left[Bias\left(\hat{\theta}\right)\right]^2 \tag{10}$$

For the suitability of comparisons, denote $MSE\left(\hat{\beta}_{CMLE}\right) = Q\Lambda^{-1}Q'$, $\Lambda = diag\left(\lambda_1, \dots, \lambda_p\right) = Q'\left(X'\widehat{\Psi}_{LP}X\right)Q$, where $\lambda_1 > \lambda_2 > \cdots \lambda_p > 0$ are ordered eigenvalues of $\left(X'\widehat{\Psi}_{LP}X\right)$ matrix, Q is the orthogonal matrix whose columns constitute the eigenvectors of $\left(X'\widehat{\Psi}_{LP}X\right)$, and the j^{th} element of $Q'\beta$ is denoted as ω.

- The bias of $\hat{\beta}_{RCML}$ is

$$Bias(\hat{\beta}_{RCML}) = -k\sum_{j=1}^{p}\frac{\omega_j}{(\lambda_j+k)} \tag{11}$$

- The variance of $\hat{\beta}_{RCML}$ is

$$Var(\hat{\beta}_{RCML}) = \sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2} \tag{12}$$

- Then the TMSE of $\hat{\beta}_{RCML}$ is

$$TMSE(\hat{\beta}_{RCML}) = \sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2} + \sum_{j=1}^{p}\frac{k^2\omega_j^2}{(\lambda_j+k)^2} \tag{13}$$

3. ESTIMATION OF THE RIDGE PARAMETER

The major objective of logistic panel ridge regression method is to find a suitable k to the extent that the decrease in variance of the ridge regression estimator assures the increase in its bias. The first method of choosing the biasing or ridge parameter k was proposed by Hoerl and Kennard [10] for the linear regression model. It states that there always exists a k > 0. Schaefer et al. [12] followed the same principal to find the ridge parameter for logistic regression. To show that we minimize the mean squared error of ridge estimator, the first derivative of Eq. (13) with respect to k is

$$\frac{\partial\, TMSE(\hat{\beta}_{RCML})}{\partial k} = -2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^3} + 2k\sum_{j=1}^{p}\frac{k\omega_j^2}{(\lambda_j+k)^3} \tag{14}$$

Since $\lambda_j > 0$, the first derivatives of the first and second terms are always non-positive and non-negative, respectively. Moreover, the optimal value of any individual parameter k_j can be found by setting Eq. (14) to zero and solve for *k*. At that time, it can be illustrated as follows:

$$k_{opt} = k_j = \frac{1}{\omega_j^2} \tag{15}$$

The optimal value of k_j fully depends on the unknown ω_j, which can be estimated from the data. As Hoerl and Kennard [10], Schaefer et al. [12] proposed replacing ω_j by its estimator $\widehat{\omega}_j$. That means the optimum value if *k* is obtained as follows:

$$\hat{k}_{opt} = \widehat{k}_j = \frac{1}{\widehat{\omega}_j^2} \tag{16}$$

- Following Hoerl and Kennard [10] and Hoerl et al. [11], we can use the following estimators

for *k*:

$$\hat{k}_1 = \frac{1}{\hat{\omega}^2_{max}} \tag{17}$$

$$\hat{k}_2 = \frac{p}{\sum_{j=1}^{p} \hat{\omega}_j^2} \tag{18}$$

- Following Kibria [25], we can use the following estimators for *k*:

$$\hat{k}_3 = \frac{1}{\left(\prod_{j=1}^{p} \hat{\omega}_j^2\right)^{1/p}}; \quad \hat{k}_4 = median\left(\frac{1}{\hat{\omega}_j^2}\right) \tag{19}$$

- Following Muniz et al [26], we can use the following estimator for *k* based on the square root of the geometric mean of $\hat{k}_j$:

$$\hat{k}_5 = \left(\prod_{j=1}^{p} \sqrt{\frac{1}{\hat{\omega}_j^2}}\right)^{\frac{1}{p}} \tag{20}$$

- Following Kibria et al. [27], we can use the following estimator for *k*:

$$\hat{k}_6 = median\left(\frac{\lambda_j}{(NT-p)+\lambda_j \hat{\omega}_j^2}\right) \tag{21}$$

- Following Dorugade [28], we can use the following estimator for *k*:

$$\hat{k}_7 = \frac{2}{\lambda_{max} \hat{\omega}_j^2} \tag{22}$$

- Following the previous works, we propose the following estimators for *k*:

$$\hat{k}_{new1} = mean\left(\frac{1}{\hat{\omega}_j^2}\right)^{\frac{1}{p}} \tag{23}$$

$$\hat{k}_{new2} = \left(\prod_{j=1}^{p} \left(\frac{1}{(1+\hat{k}_1)\hat{\omega}_j^2}\right)\right)^{\frac{1}{p}} \tag{24}$$

4. EMPIRICAL STUDY

Weekly data for sixty-seven female patients who received adjuvant chemotherapy following mastectomy for Breast Cancers are utilized during the three-month period in a hospital in Port Said, Egypt, where patients' data are recorded while they are visiting the hospital for diagnosis and treatment [9]. We employ longitudinal data composed of repeated measurements, where response variable are assessed at multiple time points for each patient in which responses are binary

response variable (Take chemo or no) coding either no delayed completion (1 = complete) or delayed completion (0 = no complete) of adjuvant chemotherapy in patients with Breast Cancer. There are changes between patients and over time under the influence of determining factors (Age, BSA, HT, WT, HGB, WBC, GRAN, ALT, AST, WBC, PLT, SRCR, RBC and Urea). Variations between patients were allowed, as there is a protocol for distributing sessions for each patient (six sessions), but a specific schedule for treatment within the hospital must be followed which is three months because, after three months, the patient's condition is re-evaluated to determine the patient's response to chemotherapy treatment (based on the results of tumor evidence analysis).

We employ one specification of logistic panel data models when the individual effects are fixed, but this data has multicollinrarity problem, the fact that justifies using ridge estimators to solve this problem and to choose the appropriate model for our data.

4.1. Data Description

As an empirical application, this paper is concerned with studying the most vital factors that affect delayed completion of adjuvant chemotherapy in patients caught with Breast Cancer using data for sixty-seven patients during a three-month period. The available data set is restricted to the amount of information available for each patient involved. We selected a variety of explanatory variables that have been shown to correlate with the adjuvant chemotherapy available for patients diagnosed with Breast Cancer. It should be noted that the data were used in another paper by the same researchers [9], but new variables have been added. These variables were related to each other. It caused a multicollinearity problem.

Table1. Definition of the Variables

Variable	Definition
Response: y	The doctor's decision to give the adjuvant chemotherapy session to the patient: Take Y_{it}=1) or no Adjuvant (Y_{it}=0) Chemotherapy. The count of ones in Y is 290 (i.e., it present 72% of the sample).
Covariates	
Age	The ages of Breast Cancer patients who were included in the study (67patients) ranged from 25 to 81 years and in this study the patients (females only).
HT	Length
WT	Weight
BSA	Body surface area
HGB	Haemoglobin
WBC	White blood cell count
PLT	Platelets
GRAN	Granulocytes count
ALT	Alanine transaminase
AST	Aspartate aminotransferase
SRCR	Creatinine
Urea	Blood Urea Nitrogen
LYM	Lymphocytes count
RBC	Red blood cell count

4.2. Descriptive Statistics

The used software in our study is "R version 4.0.1". Table 1 displays the definition of the used variables, and some descriptive statistics of these variables have been presented in the Table 2 to show the Age of breast cancer patients ranged from 25 to 76 years mean 51.7 with standard deviation 12.7, Haemoglobin (HGB) ranged from (6 to 15.2) with mean 10.13; standard deviation 1.56, Platelets (PLT) ranging from 54 to 556 with mean 236.9 and standard deviation (SD) 86.9, Alanine transaminase (ALT) ranging from .7 to 90 with mean 21.4 and standard deviation 11.6. In general, since the coefficient of variation (CV) values of all variables are less than 1, this means that the data do not have large variation, and then we do not expect outlier values in the data. It show that the CV of all variables less than one, then the data not have large variation.

Table.2 Descriptive Statistics of the Variables

Variable	Mean	SD	CV	Min.	Max.
y	0.7	.45	0.64	0	1
Age	51.7	12.7	0.24	25	76
HT	157.7	12.5	0.08	100	176
WT	80.5	16.4	0.20	42	120
BSA	1.8	.167	0.09	1.3	2
HGB	10.1	1.56	0.15	6	15.2
WBC	6.5	3.62	0.55	1.8	57
PLT	236.9	86.9	0.36	54	556
GRAN	2.7	1.2	0.44	.3	8.2
ALT	21.4	11.64	0.55	.7	90
AST	22.8	9.74	0.43	6	78
SRCR	0.9	.84	0.93	.32	17
Urea	27.5	12.6	0.45	4	86
LYM	1.6	.99	0.62	.45	14.25
RBC	4.4	.39	0.08	3.8	6.00

4.3. Testing the Multicollinearity

The first step of data processing is to try to ensure that there is no high linear correlation between two or more explanatory variables. Statistical inferences are not reliable in the case of multicollinearity because it makes estimates of the regression coefficients inaccurate, inflates their standard errors, deflates the partial t-tests for them, gives false non-significant p-values, and reduces the predictability of the model, see [1]. We use the most common methods to detect multicollinearity:

- Pearson correlation matrix between each pair of predictor variables:

Table 3 shows that there is a strong correlation among the variables: WT (weight) with BSA (Body surface area) =0.9, HGB (Haemoglobin) with RBC (Red blood cell count) =0.9, WBC (White blood cell count) with LYM (Lymphocytes count) = 0.9, and ALT (Alanine transaminase) with AST (Aspartate aminotransferase) = 0.8

- Variance Inflation Factor (VIF) based on the results of the FELPD model:

The results of VIF with all regressors confirmed that there is multicollinearity problem among the regressors, in which, as common in most of empirical studies, the general rule of thumb lurks in the fact that VIF values more than 4 or 5 need more investigation. On the other hand, VIF values more than 10 confirm serious multicollinearity that requires correction. According to Paul [29] the results of table 3 refer to the fact that the data available are of a multicollinearity problem due to the variable (WT, BSA, HGB, WBC, LYM, RBC) value of VIF of more than 4. To investigate the presence of multicollinearity covariates variables are given in table 3.

Table 3: Correlation analysis of explanatory variables

	Age	**HT**	**WT**	**BSA**	**HGB**	**WBC**	**PLT**	**GRAN**	**ALT**	**AST**	**SRCR**	**Urea**	**LYM**	**RBC**
Age	1													
HT	.046	1												
WT	.29**	.03	1											
BSA	.3**	.4**	**.9****	1										
HGB	.042	-.03	-.02	-.03	1									
WBC	.12*	-.11*	-.02	-.06	.2**	1								
PLT	-.2**	-.16**	-.03	-.05	.2**	.62	1							
GRAN	.16**	.02	.25**	.23**	.41**	.16**	.103*	1						
ALT	-.012	.07	.3	.05	.22**	.03	.14**	.103*	1					
AST	.11*	.023	.06	.03	.09*	.012	.03	.07	**.8****	1				
SRCR	.07	.019	-.002	-.009	.04	.05	0	.08	.01	.03	1			
Urea	.09	.027**	-.07	-.2**	.01	.23**	-.12*	.09	-.09	-.06	.008	1		
LYM	.08	-.1	.04	0	.2**	**.9****	.08	.2**	.1*	.08	.04	.2**	1	
RBC	.04	.16**	.094	.15**	**.9****	.05	*.12*	.102*	.026	-.05	.04	.19*	.02	1
VIF	**1.35**	**9.45**	**2.61**	**11.1**	**7.79**	**4.65**	**1.25**	**2.9**	**2.53**	**1.3**	**1.15**	**1.39**	**4.66**	**7.52**

Notes: The superscripts ** and * indicate statistical significance at the 0.001 and 0.01 level, respectively.

4.4. Results of Different Models

In case of logistic panel analysis, response variable is binary response (Take or not a chemo session) variation under the influence of determinant factors. During the three-month period of observation, there are fourteen medical factors of different effects for each patient.

Table 4 presents the result of pooled logit model versus fixed effect logit models, Hausman Test, and Akaike's information criterion (AIC).

Table 4. Results of Logit Panel Data Models

Variable	Pooled Logit Model	Fixed Effects Logit Models	
	UPL	UCML	CML
Intercept	-7.299**	-6.89**	---------
Age	0.01	.01	.01
WT	-0.01	-.01	.01
HT	-0.08	-.01	.01
BSA	3.27	2.73	.92
HGB	0.44	.5*	.26
WBC	-0.03	-.01	.02
PLT	0.02	.001	.003*
ALT	-0.06***	-.06***	-.06**
AST	-0.01	-.004	.01
GRAN	0.115	.12	.21
SRCR	-0.64	-.62	-.76
Urea	0.01	.02	.01
LYM	-0.02	-.03	-.07
RBC	0.18	.08	.28
Goodness of Fit			
AIC	428.83	427.8	379.8
Hausman test	----------	$\chi^2 = 32.304$, df = 12, P-value = .0012	

Notes: The superscripts ***, **, and * indicate statistical significance at the 0.001, 0.01 and 0.05 level, respectively.

Table 4 summarizes the results of the three estimations. The results indicate that the unconditional pooled logit (UPL) and UCML estimates are roughly similar; where the ALT variable is significantly negative in both estimations, BSA and HGB variables are significantly positive in both estimations. However, the AIC value of UCML estimation is smaller than the AIC value of UPL estimation, and then the UCML estimation is better than UPL estimation. While in CML estimation, the significant variables (PLT and ALT) are different from those in UPL and UCML are estimations namely (HGB and ALT). To choose the best estimation of this data, we used AIC and Hausman's specification test. Since, the p-value of the Hausman's specification test is less than 0.05, then we can reject the null hypothesis of this test. This means that CML estimate is more consistent and efficient than the UCML estimate. This conclusion is confirmed by AIC, where the CML estimate has the smallest values of AIC.

Table 5. The Coefficient Estimates, AIC, and MSE of CML and RCML Estimators

Variable	CML	RCML									
		$\widehat{k}_{opt}$	$\widehat{k}_1$	$\widehat{k}_2$	$\widehat{k}_3$	$\widehat{k}_4$	$\widehat{k}_5$	$\widehat{k}_6$	$\widehat{k}_7$	$\widehat{k}_{new1}$	$\widehat{k}_{new2}$
Age	.01	.01	.01	.05	.03	.06	.02	.06	.06	.06	.06
WT	.01	-.03	-.03	-.07	.08	.06	.08	-.02	.01	-.08	-.01
HT	.01	-.09	-.09	-.07	-.02	-.04	-.02	-.08	-.07	-.02	-.08
BSA	.92	.33	.32*	.49	.15**	.21*	.13**	.31*	.29*	.39	.31*
HGB	.26	.06***	.06***	.08**	.03***	.05***	.03***	.06***	.06***	.07**	.06***
WBC	.02	.08	.08	.01	.01	.01	. 01	.08	.09	.05	.08
PLT	.03*	.03	.03	.03	.02	.02	.02	.03	.03	.03	.03
ALT	-.06**	-.08***	-.08***	-.09***	-.05***	-.06***	-.05***	-.08***	-.08***	-.09***	-.08***
AST	.01	-.03	-.03	-0.3	-.04***	-.04*	-.04***	-.03	-.04	-.03	-.04
GRAN	.21	.02	.01	.02	.02	.02	.01*	.02	.02	.02	.02
SRCR	-.76	-.04*	-.04*	-.05	-.02*	-.03*	-.06*	-.03*	-.04*	-.05*	-.04*
Urea	.01	.04	.04	.04	.05	.05	.05	.04	.04	0.04	.04
LYM	-.07	-.08	-.08	-.06	-.04	-.07	-.03	-.08	-.01	-.01	-.01
RBC	.28	.03	.04	.01	.01***	.05*	.04***	.03	.04	.03	.04
MSE	**31.01**	**5.89**	**5.95**	**7.71**	**9.03**	**7.16**	**9.64**	**5.91**	**6.04**	**6.17**	**5.94**
AIC	**379.8**	**-700.5**	**-700.63**	**-698.14**	**-700.15**	**-702.09**	**-698.57**	**-700.81**	**-701.23**	**-699.62**	**-700.93**

Notes: The superscripts ***, **, and * indicate statistical significance at the 0.001, 0.01 and 0.05 level, respectively.

Table 5 presents the results of CML and RCML estimators. Table 5 summarizes the estimated predicted probability along with the ridge regression coefficients of the estimators that are presented and to provide the p-values for testing the significance of regression parameters of the CML and RCML estimators. We can show that the ALT and PLT variables are significant, but the model has high multicollinearity because coefficients can have high standard errors and low significance even though they may be jointly significant. As a solution to this problem, use the RCML estimator. From Table 5, we observed that all ridge regression estimators have minimum MSE and AIC than that of the CML estimator. Also, proposed k_{opt} estimators have worked out well if compared to other k estimators. Then the most prominent variables that influenced delayed completion of adjuvant chemotherapy in patients diagnosed with breast cancer chemotherapy after mastectomy are Haemoglobin (HGB), Alanine transaminase (ALT) and Creatinine (SRCR).

5. CONCLUSIONS

In this paper, we presented a new estimator that was used for empirical study to be used in further studies. An empirical study was done to figure out the most significant factors that influence delayed completion of adjuvant chemotherapy in patients diagnosed with breast cancer plus adjuvant chemotherapy outcomes in patients who caught breast cancer to shed light on the link between chemotherapy duration and its existing outcomes, in this regard.

To achieve the paper goal, we used two estimation methods of fixed effects logistic panel model: fixed effects by unconditional maximum likelihood (UCML) and conditional maximum likelihood (CML) estimators. The study results show that the CML estimator is efficient and better than the UCML estimator. However, the resulting model have high multicollinearity. Therefore, we proposed the ridge conditional maximum likelihood (RCML) estimator for this model and compared the CML estimator with the proposed estimator. The results showed that the RCML estimator was superior to the CML estimator in the sense of smaller MSE, and it is evident that it is very critical to use the RCML estimator over CML estimator when explanatory variables are correlated. Furthermore, the outcomes manifest that the most significant effects that influence

delayed completion of adjuvant chemotherapy in those patients receiving breast cancer chemotherapy after mastectomy are Body Surface Area (BSA), Haemoglobin (HGB), Alanine Transaminase (ALT) and Creatinine (SRCR).

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] A.H. Studenmund, Using econometrics: a practical guide, Seventh Edition, Pearson, Boston, (2017).

[2] B. Baltagi, Econometric analysis of panel data, Wiley, New York, (2008).

[3] M.R. Abonazel, Different estimators for stochastic parameter panel data models with serially correlated errors, J. Stat. Appl. Probab. 7 (2018), 423-434. https://doi.org/10.18576/jsap/070303.

[4] [1] M. Reda Abonazel, Generalized estimators of stationary random-coefficients panel data models, REVSTAT-Stat. J. 17 (2019), 493-521. https://doi.org/10.57805/REVSTAT.V17I4.278.

[5] A.H. Youssef, M.R. Abonazel, Alternative GMM estimators for first-order autoregressive panel model: An improving efficiency approach, Commun. Stat. – Simul. Comput. 46 (2016), 3112–3128. https://doi.org/10.1080/03610918.2015.1073307.

[6] M.R. Abonazel, O.A. Shalaby, Using dynamic panel data modeling to study net FDI inflows in MENA countries, Stud. Econ. Econ. 44 (2020), 1–28. https://doi.org/10.1080/10800379.2020.12097360.

[7] M.R. Abonazel, O. Shalaby, On labor productivity in OECD countries: Panel data modeling, WSEAS Trans. Bus. Econ. 18 (2021), 1474–1488. https://doi.org/10.37394/23207.2021.18.135.

[8] A.H. Youssef, M.R. Abonazel, O.A. Shalaby, Spatial and non-spatial panel data estimators: Simulation study and application to personal income in U.S. States, WSEAS Trans. Math. 21 (2022), 487–514. https://doi.org/10.37394/23206.2022.21.56.

[9] A.M. El-Masry, A.H. Youssef, M.R. Abonazel, Using logit panel data modeling to study important factors affecting delayed completion of adjuvant chemotherapy for breast cancer patients, Commun. Math. Biol. Neurosci. 2021 (2021), 48. https://doi.org/10.28919/cmbn/5410.

[10] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics, 12 (1970), 55–67.

[11] A.E. Hoerl, R.W. Kannard, K.F. Baldwin, Ridge regression: some simulations, Commun. Stat. 4 (1975), 105–123. https://doi.org/10.1080/03610927508827232.

[12] R.L. Schaefer, L.D. Roi, R.A. Wolfe, A ridge logistic estimator, Commun. Stat. – Theory Methods. 13 (1984), 99–113. https://doi.org/10.1080/03610928408828664.

[13] F.A. Awwad, K.A. Odeniyi, I. Dawoud, et al. New two-parameter estimators for the logistic regression model with multicollinearity, WSEAS Trans. Math. 21 (2022), 403–414. https://doi.org/10.37394/23206.2022.21.48.

[14] M.R. Abonazel, R.A. Farghali, Liu-type multinomial logistic estimator, Sankhya B. 81 (2018), 203–225. https://doi.org/10.1007/s13571-018-0171-4.

[15] A.F. Lukman, B. Aladeitan, K. Ayinde, et al. Modified ridge-type for the Poisson regression model: simulation and application, J. Appl. Stat. 49 (2021), 2124–2136. https://doi.org/10.1080/02664763.2021.1889998.

[16] M.N. Akram, M.R. Abonazel, M. Amin, et al. A new Stein estimator for the zero‐inflated negative binomial regression model, Concurrency Comput. 34 (2022), e7045. https://doi.org/10.1002/cpe.7045.

[17] R. Nissen-Meyer, K. Kjellgren, K. Malmio, et al. Surgical adjuvant chemotherapy.Results with one short course with cyclophosphamide after mastectomy for breast cancer, Cancer. 41 (1978), 2088–2098. https://doi.org/10.1002/1097-0142(197806)41:6<2088::aid-cncr2820410604>3.0.co;2-j.

[18] E.M.A. Bleiker, H.M. van der Ploeg, J.H.C.L. Hendriks, et al. Personality factors and breast cancer development: a prospective longitudinal study, J. Nat. Cancer Inst. 88 (1996), 1478–1482. https://doi.org/10.1093/jnci/88.20.1478.

[19] M. Chavez-MacGregor, C.A. Clarke, D.Y. Lichtensztajn, et al. Delayed initiation of adjuvant chemotherapy among patients with breast cancer, JAMA Oncol. 2 (2016), 322-329. https://doi.org/10.1001/jamaoncol.2015.3856.

[20] F. Bray, J. Ferlay, I. Soerjomataram, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians. 68 (2018), 394–424. https://doi.org/10.3322/caac.21492.

[21] W. Greene, The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects, Econometrics J. 7 (2004), 98–119. https://doi.org/10.1111/j.1368-423x.2004.00123.x.

[22] G. Chamberlain, Analysis of covariance with qualitative data, Rev. Econ. Stud. 47 (1980), 225-238. https://doi.org/10.2307/2297110.

[23] E.B. Andersen, Asymptotic properties of conditional maximum-likelihood estimators, J. R. Stat. Soc. Ser. B (Methodol.) 32 (1970), 283–301. https://doi.org/10.1111/j.2517-6161.1970.tb00842.x.

[24] J.S. Kunz, K.E. Staub, R. Winkelmann, Estimating fixed effects: perfect prediction and bias in binary response panel models, with an application to the hospital readmissions reduction program, SSRN. (2017). https://doi.org/10.2139/ssrn.3074193.

[25] B.M.G. Kibria, Performance of some new ridge regression estimators, Commun. Stat. – Simul. Comput. 32 (2003), 419–435. https://doi.org/10.1081/sac-120017499.

[26] G. Muniz, B.M. Kibria, G. Shukur, On developing ridge regression parameters graphical investigation, J. Int. Stat. 39 (2012),115-138.

[27] B.M.G. Kibria, K. Månsson, G. Shukur, Performance of some logistic ridge regression estimators, Comput. Econ. 40 (2011), 401–414. https://doi.org/10.1007/s10614-011-9275-x.

[28] A.V. Dorugade, New ridge parameters for ridge regression, J. Assoc. Arab Univ. Basic Appl. Sci. 15 (2014), 94–99. https://doi.org/10.1016/j.jaubas.2013.03.005.

[29] R.K. Paul, Multicollinearity: causes, effects and remedies, Indian Agricultural Research Institute, New Delhi, (2006).