Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2021, 2021:48
https://doi.org/10.28919/cmbn/5410
ISSN: 2052-2541

USING LOGIT PANEL DATA MODELING TO STUDY IMPORTANT FACTORS AFFECTING DELAYED COMPLETION OF ADJUVANT CHEMOTHERAPY FOR BREAST CANCER PATIENTS

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: In analysing panel data in which the dependent variable is a binary choice variable taking values 1 or 0 for success or failure respectively, it is feasible to consider the conditional probabilities of the dependent variable. Under strict exogeneity, this conditional probability equals the expected value of the dependent variable. This treatment calls for a nonlinear function which will ensure that the conditional probability lies between 0 and 1, and such functions yield the probit model and the logit model. This paper discusses an estimation of nonlinear logit panel data model with fixed effects. There are two main estimators for such models: 'unconditional maximum likelihood' and 'conditional maximum likelihood'. Application study was designed to determine the most important factors affecting delayed completion of adjuvant chemotherapy among patients with breast cancer and adjuvant chemotherapy improvement outcomes of patients with breast cancer to determine the relationship between time to chemotherapy and outcome according to breast cancer. The optimal timing from beginning to the end of chemotherapy is known (three months). We hypothesized that prolonged time to chemotherapy would be associated with adverse outcomes. Delayed time to chemotherapy was defined as more three months from the first dose and the

*Corresponding author
E-mail address: mabonazel@cu.edu.eg
Received January 9, 2021

last dose of chemotherapy. The study results show that the conditional fixed effects logit estimator is efficient and better than the unconditional pooling and unconditional fixed effects logit estimators. And we find that the most important factors affecting delayed completion of adjuvant chemotherapy among patients are haemoglobin, platelets, and alanine transaminase.

Keywords: panel data; logistic regression; fixed effects, maximum likelihood estimator; conditional maximum likelihood estimator.

2010 Subject Classification: 62J12, 62P10.

1. INTRODUCTION

The panel or longitudinal data are becoming increasingly popular in applied work, as they offer several advantages over pure cross-sectional or pure time-series data. A particularly useful feature is that they allow researchers to use the techniques of panel data estimation that can take such heterogeneity explicitly into account by allowing for subject-specific variables. In this regard, it helps to boost sample size which makes the results more reliable. It increases the number of data points, degree of freedom and reduced multicollinearity of explanatory variables. The panel data are better suited to study the dynamics of change, typically assumed to be linear, and that captures the statistical relationship between the dependent and the explanatory variables. see, e.g., [1, 2, 3, 4, 5] for more details on panel data modeling.

Logistic regression analysis is the most popular regression technique available for modeling dichotomous dependent variables. Logistic regression can describe the relationship between a categorical outcome and a set of covariates (predictor variables). The categorical outcome may be binary. For example, studies determining the most important factors affecting delayed completion of adjuvant chemotherapy among patients with breast cancer, the take chemo (1) or do not take chemo (0). The predictor variable(s) may be continuous or categorical, and an advantage of using Logistic Regression is that it allows properties of a linear regression model to be exploited. The logit itself can take values between $-\infty$ and ∞, and probability remains constrained between 0 and 1. The candidate predictor variables do not have to be (Normally distributed - Linearly related - Have equal variances), and candidate predictor variables can be (continuous, dichotomous). See, e.g., [6, 7, 8, 9] for more details on logistic regression modeling.

In the analysis of panel data, the dependent variable is a binary choice variable taking values 1 or 0 for success or failure, respectively. This treatment calls for a nonlinear panel models requiring conditioning the probabilities on the minimum sufficient statistic for the fixed effects so as to curb the incidental parameter problem. Parametric logistic panel data models with individual time-invariant effects can be estimated by either the conditional or the unconditional fixed effects maximum likelihood estimator.

Breast cancer is the most common 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. In recent years, cancer rates have been shown to be rising steadily in low-and middle-income countries (according to the WHO global burden of disease report).

Chemotherapy for Breast Cancer uses anti-cancer drugs that may be given intravenously (injected into your vein) or orally. The drugs travel through the bloodstream to reach cancer cells in most parts of the body. Chemotherapy drugs used for breast cancer can be given 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 might be given to try to shrink the tumor, so it can be removed with less extensive surgery.

Because of this, neoadjuvant chemo is often used to treat cancers that are too big to be removed by surgery when first diagnosed (called locally advanced cancers). For certain types of breast cancer, if there are tumor cells still found at the time of surgery (also called residual disease), Many patients may be offered more chemotherapy after surgery to reduce the chances of the cancer not to come back (recurrence). 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.

This paper is organized as follows: Section 2 provides a review of the literature on this topic. Section 3 provides the logit panel data model with fixed effects and the estimation methods. The results of the empirical study are presented in Section 4. Section 5 offers some concluding remarks.

2. REVIEW OF LITERATURE

Meyer et al. [10] 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. The result: the control group now has 234 recurrences and 196 deaths, and the treatment group 175 recurrences and 146 deaths. The differences of fifty deaths in favour of the treatment group are significant. The differences in recurrence rates increased gradually, to reach 10.71% four years after mastectomy, and to remain at the same level for another six years. The differences in death rates increased until six years after mastectomy to be 10.48% after 10 years. With this pattern, the mechanism is probably not a delay in onset of clinical recurrences, but a definite reduction of recurrence rates due to tumoricidal chemotherapy. The same chemotherapy course given three weeks after mastectomy seemed without effect.

Eveline et al. [11] 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 the development of breast cancer. They used longitudinal study design from 1989 through 1990, a personality questionnaire was sent to all female residents of the Dutch city of Nijmegen who were forty-three years of age or older. In

investigation the extent to which personality factors, in addition to somatic risk factors, he may be associated with the development of primary breast cancer. Results: Personality questionnaires were sent to 28940 women and 9705 (34%) returned in such a way that they could be used for statistical analyses. Three variables were found to be statistically significantly associated with an increased risk of breast cancer: 1) having a first-degree family member with breast cancer, 2) nulliparity and 3) a relatively high score on the personality scale of anti-emotionality.

Mariana et al. [12] used logistic regression and Cox proportional hazard models to identify the determinants in delayed chemotherapy initiation of adjuvant chemotherapy among Patients with Breast Cancer and to determine the relationship between time to chemotherapy (TTC) and outcome according to breast cancer subtype. Using data from the California Cancer Registry, we studied a total of 24843 patients with breast cancer diagnosed between 2005 and 2010 and treated with adjuvant chemotherapy. Results factors associated with delays in TTC included low socioeconomic status, nonprivate insurance and Hispanic ethnicity or Non-Hispanic black race. Compared with patients receiving chemotherapy within thirty-one days from surgery, there was no evidence of adverse outcomes among those with TTC of thirty-one to sixty or sixty to ninety days. Patients treated ninety-one or more days from surgery experienced worse overall survival.

Bray et al. [13] provided a status report on the global burden of cancer worldwide using the GLOBOCAN 2018 estimates of cancer incidence and mortality produced by the International Agency for Research on Cancer, with a focus on geographic variability across twenty world regions. 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. In both sexes combined, lung cancer is the most commonly diagnosed cancer (11.6% of the total cases) and the leading cause of cancer deaths (18.4% of the total cancer deaths), closely followed by female breast cancer (11.6%), prostate cancer (7.1%), and colorectal cancer (6.1%) for incidence and colorectal cancer (9.2%), stomach cancer (8.2%), and liver cancer (8.2%) for mortality.

The global initiative for cancer registry development is an international partnership that supports better estimation, as well as the collection and use of local data, to prioritize and evaluate national cancer control efforts.

3. METHODOLOGY

3.1 Fixed Effects Logit Panel Data Model

In many studies, the dependent 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:

$$pr(y_{it} = 1|\alpha_i, \beta) = G(z) = G(\alpha_i + x_{it}\beta) \quad (1)$$

$$pr(y_{it} = 0|\alpha_i, \beta) = 1 - G(\alpha_i + x_{it}\beta) \quad (2)$$

where G(z) is a nonlinear function taking on values strictly between zero and one: $0 < G(z) < 1$, for all real number z (individual and time), 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 and the standard normal distribution, yielding the probit model. The logit panel data model takes the following form [14]:

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

where y_{it} is the response, and in the case of a logistic panel model, a binary response variable 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 cumulative distribution function (CDF) for a logistic variable, where x_{it} is $(1 \times k)$ vector of observed explanatory variables, β is $(k \times 1)$ vector of parameter or common parameters, α_i is an unobserved time invariant individual effect.

3.2 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 explanatory variables are not linear combinations of each other.

Assumption 4: The explanatory 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 Assumption 3.

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

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

3.3 Incidental Parameter Problem

For fixed effects logistic panel data (FELPD) model, the presence of individual effect complicates the parameter estimation significantly. Consider the FELPD model in equation (3), in this case α_i and β are unknown parameters to be estimated and as $N \to \infty$ for fixed T. The number of parameters α_i increases with N. As such α_i cannot be consistently estimated for fixed T. This is known as the incidental parameter problem in statistics. Moreover, if we attempt to estimate α_i directly by adding $N - 1$ individual dummy variables to logit, this will result in severely biased and consistent estimates of β unless T is large due to the incidental parameters problem. For linear panel data regression model, it is easy to eliminate α_i by means of first differencing [15].

3.4 Unconditional Fixed Effects Logit Estimator

According to the unconditional fixed effects logit (UFEL) estimator, for estimating the parameters of this model is to include dummy variables for the units and then to maximize the unconditional likelihood function. The fixed effects model is specified by including a separate fixed effect for each identified panel or group in the data. Standard logistic regression commands

can be used to estimate the fixed effects model. The unconditional maximum likelihood estimator (MLE) of parameters is consistent as $T \to \infty$ for fixed N but inconsistent as $N \to \infty$ for fixed T. The binary y_{it} are assumed to be independent both between and within groups with

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

If $y_{it} = 1$ for all t then the MLE of α_i is ∞, $y_{it} = 0$ for all t then the MLE of α_i is $-\infty$.

The unconditional log-likelihood function of the FELPD model can be expressed as follows [16]:

$$lnL = \sum_{i=1}^{N} \sum_{t=1}^{T} \{y_{it} lnG(\alpha_i + x_{it}\beta) + (1 - y_{it}) ln[1 - G(\alpha_i + x_{it}\beta)]\} \quad (4)$$

The MLEs for β and α are given by: $\frac{d\, lnL}{d\, \beta} = 0$, $\frac{d\, lnL}{d\, \alpha} = 0$, where $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$; $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)'$. Generally, the fixed effects unconditional likelihood function does not have explicit solution for the parameter estimates in terms of the data and must be solved iteratively with Newton-Raphson or a scoring algorithm, see, e.g., [14, 17, 18].

3.5 Conditional Fixed Effects Logit Estimator

The conditional maximum likelihood estimator of the FELPD model is usually called the conditional fixed effects logit (CFEL) estimator. We must emphasize that the CFEL estimator does not arise by treating the α_i as parameters to be estimated along with β.

The conditional likelihood approach can be applied directly to the fixed effects logit probability 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 outcomes are (0, 1) and (1, 0) with explanatory variables. The conditional maximum likelihood estimator of β can be obtained simply from a standard maximum likelihood binary logit programme.

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 $\mathrm{y_{i1}} + \mathrm{y_{i2}} = 0$ or 2 then $\mathrm{y_{i1}}$ and $\mathrm{y_{i2}}$ are both determined given their sum. So, the only case of interest is $\mathrm{y_{i1}} + \mathrm{y_{i2}} = 1$. Then the two possibilities are $\mathrm{w_i} = 1$ if $(\mathrm{y_{i1}} = 0, \mathrm{y_{i2}} = 1) = (0,1)$ and $\mathrm{w_i} = 0$ if $(\mathrm{y_{i1}} = 1, \mathrm{y_{i2}} = 0) = (1,0)$. The conditional density is:

$$pr(w_i = 1|y_{i1} + y_{i2} = 1) = \frac{pr(w_i = 1)}{pr(w_i = 0) + pr(w_i = 1)}$$

$$= \frac{e^{\beta'(x_{i2} - x_{i1})}}{1 + e^{\beta'(x_{i2} - x_{i1})}} = G\left(\beta'(x_{i2} - x_{i1})\right)$$

which does not depend on α_i. The conditional log-likelihood function is:

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

where $I_1 = \{i|y_{i1} + y_{i2} = 1\}$. The conditional maximum likelihood estimator is a consistent estimator of β, regardless of whether α_i and x_{it} are correlated, see [19, 20].

4. EMPIRICAL STUDY

Weekly data for sixty-seven female patients treated with adjuvant chemotherapy after mastectomy for breast cancers are used during the three-month period in Al-Tadamun Hospital for the treatment of tumors in Port Said Governorate in Egypt, where patients' data are recorded while they are visiting the hospital for diagnosis and treatment.

We employ longitudinal data composed of repeated measurements, where outcomes are assessed at multiple time points for each patient in which outcomes 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. Changes between patients and over time under the influence of determining factors (Age, BSA, HT, HGB, WBC, GRAN, ALT, WBC, PLT, SRCR, RBC, Urea). Variations between patients were allowed, as there is 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 two specifications of logistic panel data models (UCLE and CLE) when the individual effects are fixed, to choose the appropriate model for our data.

4.1 Data Description

As an empirical application, this paper is concerned with studying the most important factors affecting delayed completion of adjuvant chemotherapy among patients with breast cancer using data for sixty-seven patients during a three-month period. The data set is limited by 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 among patients with breast cancer. 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 Table 2. Table 2 shows that the Age of breast cancer patients ranged from 25 to 76 years with mean 51.7 and SD 12.7. In general, since the CV values of all variables are less than 1, this means that the data do not have large variation, then we do not expect outlier values in the data.

Table 1. Definition of variables

Variable	**Definition**
Dependent: 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)
Explanatory	
Age	The age of breast cancer patient (as a demographic variable).
HT	Length
BSA	Body surface area
HGB	Haemoglobin
WBC	White blood cell count
PLT	Platelets
GRAN	Granulocytes count
ALT	Alanine transaminase
SRCR	Creatinine
Urea	Blood Urea Nitrogen
RBC	Red blood cell count

Table 2. Descriptive Statistics of the Variables

Variable	Mean	SD	CV	Min.	Max.
Age	51.70	12.70	0.25	25.00	76.00
HT	157.70	12.53	0.08	100.00	176.00
BSA	1.81	0.17	0.09	1.33	2.00
HGB	10.13	1.56	0.15	6.00	15.20
WBC	6.46	3.62	0.56	1.80	57.00
PLT	236.91	86.90	0.37	54.00	556.00
GRAN	2.68	1.18	0.44	0.30	8.20
ALT	21.37	11.64	0.54	0.70	90.00
SRCR	0.97	0.84	0.87	0.32	17.00
Urea	27.46	12.60	0.46	4.00	86.00
RBC	4.43	0.39	0.09	3.24	5.34

Notes: SD: Standard Deviation, CV: Coefficient of Variation, Min.: Minimum value, and Max.: Maximum value.

4.2 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. The first step of data processing is to try to ensure that there is no high linear correlation between two or more explanatory variables. Where 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 [21]. We use the most common methods to detect multicollinearity: (i) Pearson correlation matrix between each pair of explanatory variables and (ii) the Variance Inflation Factor (VIF). See e.g., [22, 23, 24] for handling and solving this problem in regression models if this problem exists.

Table 3. Pearson Correlation Matrix and VIF

	Age	HT	BSA	HGB	WBC	PLT	GRAN	ALT	SRCR	Urea	RBC
Age	1										
HT	-.17	1									
BSA	.23	.38	1								
HGB	.04	-.03	-.03	1							
WBC	.09	-.11	-.06	.204	1						
PLT	-0.2	-.16	-.06	.21	.06	1					
GRAN	.16	.02	.23	.41	.16	.103	1				
ALT	-.04	.07	.047	.215	.028	.143	.103	1			
SRCR	.06	.019	-.01	.04	.05	.00	.08	.01	1		
Urea	.21	-.27	-.16	.01	.23	-.12	.09	-.09	.01	1	
RBC	.16	.16	.15	.05	.05	-.12	.11	.03	.04	.19	1
VIF	**1.29**	**1.469**	**1.46**	**1.37**	**1.17**	**1.24**	**1.27**	**1.28**	**1.08**	**1.48**	**1.17**

Table 3 shows that there is not a strong correlation among the variables. Additionally, the general rule of thumb is that VIF values exceeding five need further investigation, while VIF values exceed ten indicate to serious multicollinearity requiring correction. Table 3 shows that results of VIF confirmed that there is no multicollinearity problem because all values of VIF less than five, see [25, 26, 27].

4.3 Estimation Results

In our analysis, the dependent 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 ten medical factors of different effects for each patient. Table 4 presents the results of three estimations: The unconditional pooling logit (UPL) estimation, UFEL estimation, and the CFEL estimation. The UPL is the unconditional maximum likelihood of the logit regression without fixed effects as in cross section models [28].

Table 4. Results of Logit Panel Data Models

Variable	Pooled Logit Model	Fixed Effects Logit Model	
	UPL	**UFEL**	**CFEL**
Intercept	-7.74**	-7.55**	------
Age	.01	.011	.008
HT	-.001	-.004	.002
BSA	2.00*	1.93*	1.85
HGB	.53***	.54***	.38***
WBC	-.01	-.02	.003
PLT	.002	.002	.003*
GRAN	.1	.12	.21
ALT	-.07***	-.07***	-.05***
Urea	.01	.002	.01
SRCR	-.68	-.63	-.59
RBC	.11	.17	.25
Goodness of fit			
AIC	**423.17**	**421.74**	**374.66**
Hausman test	-------	**χ^2=29.22 (df=10), p-value = .0011**	

Notes: The superscripts ***, **, and * indicate statistical significance at the 0.001, 0.01 and 0.05 level, respectively. For the UFEL estimation, coefficient estimates for the dummy variables are not reported. AIC: Akaike Information Criterion, and df: degree of freedom.

Table 4 summarizes the results of the three estimations. The results indicate that UPL and UFEL 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 UFEL estimation is smaller than the AIC value of UPL estimation, then the UFEL estimation is better than UPL estimation.

While in CFEL estimation, the significant variables (HGB, PLT, and ALT) are different from those in UPL and UFEL estimations namely (HGB, BSA, and ALT). We note that HGB and PLT variables are significant positive relationships at the .05 and .0001 level. Therefore, the

estimated coefficient of ALT variable is significantly negative as expected. Therefore, Age, HT, BSA, WBC, GRAN, Urea, SRCR, and RBC variables are insignificant.

To choose the best estimation of this data, we used AIC and Hausman's specification test. Firstly, in comparing the first two unconditional estimates (UPL and UFEL), the AIC indicates rejection of the pooled logit model in favour of the unconditional fixed effects logit model because it has the smallest values of AIC. Secondly, since the p-value of the Hausman's specification test less than 0.05, then we can reject the null hypothesis of this test, this means that CFEL estimate is consistent and efficient than the UFEL estimate. This conclusion is confirmed by AIC, where the CFEL estimate has the smallest values of AIC.

5. CONCLUSION

In this paper, we have reviewed two common methods of estimation for nonlinear (binary choice) logit panel data models. Detailed derivations of the conditional and unconditional maximum likelihood logit panel data estimators are discussed. In particular, we condition out the incidental parameters from the logit model thereby curbing the incidental parameter problem which would otherwise have made parameter estimation complicated.

An empirical study was designed to determine the most important factors affecting delayed completion of adjuvant chemotherapy among patients with breast cancer and adjuvant chemotherapy improvement outcomes of patients with breast cancer to determine the relationship between time to chemotherapy and outcome according to breast cancer.

To achieve this aim, the pooling, unconditional, conditional fixed effects logit estimators have been conducted. The results show that the best estimation of the model is the conditional fixed effects logit estimate, because it has the smallest values of AIC. This conclusion is confirmed by Hausman test. And the most important variables affecting delayed completion of adjuvant chemotherapy among patients with breast cancer chemotherapy after mastectomy for breast cancer are haemoglobin, platelets, and alanine transaminase.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

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

[2] A.H. Youssef, M.R. Abonazel, Alternative GMM estimators for first-order autoregressive panel model: an improving efficiency approach, Commun. Stat.-Simul. Comput. 46 (2017), 3112-3128.

[3] M.R. Abonazel, Different estimators for stochastic parameter panel data models with serially correlated errors, J. Stat. Appl. Probab. 7 (2018), 423-434.

[4] M.R. Abonazel, Generalized estimators of stationary random-coefficients panel data models: Asymptotic and small sample properties, Revstat Stat. J. 17 (2019), 493-521.

[5] M.R. Abonazel, O.A. Shalaby, Using dynamic panel data modeling to study net FDI inflows in MENA countries, Stud. Econ. Econometrics, 44 (2020), 1-28.

[6] M.R. Abonazel, M.G. Ibrahim, On estimation methods for binary logistic regression model with missing values, Int. J. Math. Comput. Sci. 4 (2018), 79-85.

[7] S. M. Mohamed, M. R. Abonazel, and M. G. Ghallab, A review of ten imputation methods for handling missing values in logistic regression: A medical application, J. Pure Appl. Sci. In press.

[8] E.A. Rady, M.R. Abonazel, M.H. Metawe'e, A comparison study of goodness of fit tests of logistic regression in R: simulation and application to breast cancer data. Acad. J. Appl. Math. Sci. 7 (2021), 50-59.

[9] S.M. Mohamed, M.G. Ghallab, M.R. Abonazel, Performance evaluation of imputation methods for missing data in logistic regression model: simulation and application. Working paper, Faculty of Graduate Studies for Statistical Research (FSSR), Cairo University, Egypt, (2021).

[10] R. Nissen-Meyer, K. Kjellgren, K. Malmio, B. Månsson, T. Norin, Surgical adjuvant chemotherapy. Results with one short course with cyclophosphamide after mastectomy for breast cancer, Cancer, 41 (1978), 2088-2098.

[11] E. Bleiker, H.M. van der Ploeg, J. Hendriks, H.J. Adèr, Personality factors and breast cancer development: a prospective longitudinal study, J. Nat. Cancer Inst. 88 (1996), 1478-1482.

[12] M. Chavez-MacGregor, C.A. Clarke, D.Y. Lichtensztajn, S.H. Giordano, Delayed initiation of adjuvant chemotherapy among patients with breast cancer, JAMA Oncol. 2 (2016), 322-329.

[13] F. Bray, J. Ferlay, I. Soerjomataram, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, Amer. Cancer J. 68 (2021), 394-424.

[14] 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.

[15] J. Neyman, E.L. Scott, Consistent Estimates Based on Partially Consistent Observations, Econometrica. 16 (1948) 1-32.

[16] J. M. Hilbe, Logistic regression models. Taylor & Francis Group, Boca Raton, (2009).

[17] B. Richard, A Newton Raphson method for the solution of system of equations, J. Math. Anal. Appl. 15 (1966), 243-252.

[18] G. Chamberlain, Analysis of covariance with qualitative data, Rev. Econ. Stud. 57 (1980), 225-238.

[19] E.B. Andersen, Asymptotic properties of conditional maximum likelihood estimators, J. R. Stat. Soc. 32 (1970), 283-301.

[20] 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.

[21] A.H. Studenmund, Econometrics: A practical guide. 7th Ed., Pearson, (2016).

[22] M.R. Abonazel, R.A. Farghali, Liu-type multinomial logistic estimator, Sankhya B. 81 (2019), 203-225.

[23] E.A. Rady, M.R. Abonazel, I.M. Taha, A new biased estimator for zero-inflated count regression models, J. Mod. Appl. Stat. Meth. In press.

[24] A.F. Lukman, B. Aladeitan, K. Ayinde, M.R. Abonazel, Modified ridge-type for the Poisson regression model: simulation and application, J. Appl. Stat. (2021), 1–13. https://doi.org/10.1080/02664763.2021.1889998.

[25] A.H. Youssef, M.R. Abonazel, E.G. Ahmed, Estimating the number of patents in the world using count panel data models, Asian J. Probabil. Stat. 6 (2020), 24-33.

[26] A.H. Youssef, M.R. Abonazel, O.A. Shalaby, Determinants of per capita personal income in U.S. states: spatial fixed effects panel data modeling, J. Adv. Res. Appl. Math. Stat. 5 (2020), 1-13.

[27] A.H. Youssef, E.G. Ahmed, M.R. Abonazel, The performance of count panel data estimators: a simulation study and application to patents in Arab countries, J. Math. Comput. Sci. In press.

[28] Y. Croissant, G. Millo. Panel data econometrics with R. Wiley, New York, (2019).