J.Natn.Sci.Foundation Sri Lanka 2011 **39** (1): 13-23

RESEARCH ARTICLE

Diagnostics for matched case control studies : SAS macro for Proc Logistic

S.D. Viswakula and M.R. Sooriyarachchi*

Department of Statistics, Faculty of Science, University of Colombo, Colombo 03.

Revised: 31 May 2010; Accepted: 17 September 2010

Abstract: Conditional logistic regression models have been extensively used in the field of medicine and mainly applied in matched case control studies. However, none of the major statistical packages, i.e. SAS, MINITAB, SPSS provide diagnostics to assess the goodness-of-fit of these models. In addition the freely downloadable package R provides no functions for this purpose. The objectives of this study are to review the available diagnostics for software for testing goodness-of-fit of conditional logistic regression models by the development of a computer programme and to test this programme which implements some of the reviewed methods on real data. The computer programme is implemented using Visual Basic for Applications (VBA) for Microsoft Excel and connected to the Statistical Analysis Software (SAS) version 9.1 using the Object Linking and Embedding (OLE) automation.

The software thus developed is tested on a matched case control study on endometrial cancer. A conditional logistic regression model is fitted to these data and the risk factors for endometrial cancer are identified. MINITAB and SPSS are incapable of doing conditional logistic regression. For testing goodness of the fitted model Proc Logistic in SAS is only capable of giving delta-beta plots which explain the influence of each observation on the parameters of the model. Besides, plots obtained from the developed computer programme, in addition provide information on stratum specific lack-of-fit statistics. These plots were very successful in identifying 3 outlying strata which were quite different from the other strata. In these 3 strata the case had not received estrogen whereas one or more control had received estrogen.

Keywords: Conditional logistic regression, dynamic data exchange (DDE), goodness-of-fit, matched case control studies, Object Linking and Embedding (OLE) automation.

INTRODUCTION

Logistic regression is widely used in both types of observational studies (prospective and retrospective). According to previous studies (Schlesselman,1982; Collett, 2003), logistic models have been extensively used in the field of medicine, and mainly applied in matched case control studies. These studies are also known as retrospective studies, where individuals with a particular condition or disease (the case) are selected for comparison with a series of individuals in whom the condition or disease is absent (the control).

Cases and controls are matched on the basis of confounding variables to control the effect of the confounding variables and this enables increase in efficiency. Matching is generally done on the basis of particular confounding variables such as age and ethnic group. In many real world situations, case control studies are used to investigate a combination of factors causing many diseases.

In conditional logistic regression models, the parameters associated with the covariates used for matching are commonly eliminated from the analysis by conditioning on the sufficient statistics of these parameters. The parameters for the uncontrolled risk factors are then estimated by maximizing the conditional likelihood function under certain assumptions. This conditioning prior to estimation tends to reduce the bias and increase the efficiency of these estimates only if the assumptions made are valid.

Checking the validity of the assumptions and the assessment of the fit of any model is important before the inferences are made. Residual analysis is vital in a conditional logistic model applied to a matched case control study, since it helps to identify matched sets which are not well fitted by the model. Such sets can have a large influence on the estimated coefficients and summary measures used to make inferences concerning various hypotheses about these parameters.

*Corresponding author (roshini@mail.cmb.ac.lk)

Assessment of the adequacy of these assumptions is seldom undertaken in practice. The lack of proper tools in frequently used statistical packages (i.e. SAS®, MINITAB, SPSS) is largely responsible for this neglect. Hence development of such tools and guidance on its usage is important. While MINITAB and SPSS provide no facilities for conditional logistic regression, SAS and EGRET have facilities for conditional logistic regression but except for delta-beta plots, no other goodness-of-fit statistics are given in either package. SAS by far is the more user-friendly compared to EGRET, thus development of this tool is considered for SAS.

The objectives of this paper are to review the literature on testing the goodness-of-fit of the conditional regression model and to develop a computer programme implementing those available methodologies and illustrate the methodologies using a matched case control data set.

The conditional logistic regression model for matched case control studies

Until recently, large multivariate studies of disease incidence have been analyzed as a series of bivariate tables. This approach can lead to over-interpretation of individual findings due to the fact that a multiplicity of tables cannot adequately represent complex interactions between the many variables under study. To help alleviate these problems, logistic regression models were introduced (Breshlow & Day, 1980) for use in prospective and retrospective studies. When the matching is done, conditional logistic regression is used to adjust for the effects of matching variables. The use of intrinsically nonlinear models is increasing in epidemiology, especially in the analysis of case-control studies by various extensions of the multiple logistic model (Moolgavkar *et al*., 1984). The parameter estimation of conditional logistic regression models for matched case control studies using the conditional likelihood has also been explained (Collett, 2003).

Diagnostics for the conditional logistic regression model

Testing goodness-of-fit: Pregibon(1984) explains a procedure for goodness-of-fit testing in conditional logistic regression models. This section summarizes this procedure.

Suppose the data consists of n matched sets and the ith set contains one case and m_i controls. Let $Y_i = (Y_{i0},...,Y_{im_i})$ be the vector of case control indicators for the ith set, where Y_{ij}=1 if the jth subject is a case and Y_{ij}=0 for a control ($0 \le j \le m_i$, $1 \le i \le n$). Also, let $\pi_{ij} = \Pr(Y_{ij}=1 / x_{ij}, \omega_i)$, where x_{ij} is a p×1 vector of covariates for the jth subject in the ith set, and ω_i is a vector of matching variables. Assume that π_{ij} follows the logistic model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha_i + \gamma' + \beta' x_{ij} . \qquad ...(1)$$

where α_i is the effect of the ith matched set, γ is the overall mean and β is the vector of parameter coefficients associated with the covariates x_{ij}.

Assuming independent responses, the appropriate conditional likelihood for inference on β is

$$L_\beta = \Pr\left(Y_1 = y_1,...,Y_n = y_n / S\right) = \frac{\exp\left(\beta' \sum_{ij} y_{ij} x_{ij}\right)}{\prod_{i=1}^{n} \sum_{h=0}^{m_i} \exp\left(\beta' x_{ih}\right)}$$

where S is the event that $\sum_j Y_{ij} = 1$, $i = 1,...,n$. The set effects α_i and γ were eliminated from the likelihood by conditioning on the number of cases per matched set. The p×1 statistic $T = \sum_{ij} Y_{ij} x_{ij}$ is sufficient for β in this conditional distribution. An alternative expression for L_β is $L_\beta = \prod_{ij} u_{ij}^{y_{ij}}$ where

$$u_{ij} = u_{ij}(\beta) = \exp\left(\beta' x_{ij}\right) / \sum_{h=0}^{m_i} \exp\left(\beta' x_{ih}\right)$$

The observed data should be arranged in a way that y_{i0}=1 and y_{i0}= 0 for $j \le 1$, $i = 1, ..., n$.

Let $\hat{\beta}$ be the conditional maximum likelihood estimator of β, that is $\hat{\beta}$ maximizes L_β, and set $u_{ij} = \hat{u}_{ij}(\hat{\beta})$. Also let $\hat{U}_i = diag\left(\hat{u}_{i0},...,\hat{u}_{im_i}\right)$ be the diagonal matrix of fitted values for the ith set and define $\hat{X}_i' = \left[\hat{x}_{i0} \mid ... \mid \hat{x}_{im_i}\right]$ to be the $p \times (m_i+1)$ matrix with columns $\hat{x}_{ij} = x_{ij} - \sum_h \hat{u}_{ih} \hat{x}_{ih}$. Large sample confidence inter vals and tests on β assume that $\hat{\beta}$ is approximately distributed as multivariate normal with mean β and covariance matrix $\left(\tilde{X}' \hat{U} \tilde{X}\right)^{-1}$, where $\hat{X}' = \left[\hat{x}_1' \mid ... \mid \hat{x}_n'\right]$ and $\hat{U} = diag(\hat{U}_1,...,\hat{U}_n)$.

The deviance for the following model $D = -2\log L_{\hat{\beta}} = -2\sum_{ij} y_{ij}\log(\hat{u}_{ij})$ is viewed as a global measure of fit over all matched sets. Suppose we are interested in checking the fit of the model to the k[th] matched set. A previous study (pregibon,1984) modeled deviations in the k[th] set from the logistic model by allowing a separate effect for each subject in this set and comparing this new model with model (1): using a score test. It was shown that the score statistic based on L_β is $s_k = z_k'\left(I - H_k\right)^{-1} z_k$ where, $z_{kj} = \left(z_{k0}, z_{k1}, ..., z_{km_k}\right)'$, is defined to be the vector of Pearson residuals for the k[th] set and $z_{kj} = \left(y_{kj} - \hat{u}_{kj}\right)/\hat{u}_{kj}^{1/2}$, I is the rank $m_k + 1$ identity matrix and $H_k = \hat{U}_k^{1/2}\tilde{X}_k\left(\tilde{X}'\hat{U}\tilde{X}\right)^{-1}\tilde{X}'_k\hat{U}_k^{1/2}$.

Pregibon(1984) approximated the null distribution of s_k with a $\chi^2_{m_k}$ distribution, where m_k is the number of controls in the k[th] matched set. Another study (Hosmer & Lemeshow, 1989) pointed out that by ignoring the off diagonal elements of the Hat matrix (H), an easily computed approximation to these statistics can be obtained, and that this approximation is accurate enough for practical purposes. The approximation is $s_k \approx \sum_j^{m_i} r_{kj}^2$, where $r_{kj} = z_{kj}\Big/\left(1-h_{kj}\right)^{1/2}$ is the standardized Pearson residual and h_{kj} is the j[th] diagonal element of H_k. The h_{kj}'s are observation leverages or potentials.

Sensitivity analysis

Influential diagnostic plots: Hosmer and Lemeshow (1989) defined the influence diagnostic as $\Delta\hat{\beta}_{kj} = \Delta X^2_{kj}\dfrac{h_{kj}}{1-h_{kj}}$ for the j[th] observation in the k[th] stratum. These values are plotted against the contribution to the likelihood as in the above case and the plot is denoted as "Individual influence statistic plot". As suggested previously (Moolgavkar *et al*., 1985; Pregibon,1984), the stratum specific total of the influence statistic is used to assess the effect of the data in an entire stratum on the fit of the model. This statistic is denoted as $\Delta\hat{\beta}_k = \sum_j^{m_i}\Delta\hat{\beta}_{kj}$ (Hosmer & Lemeshow,1989). These values are plotted against the stratum numbers to identify those strata with exceptionally large values and the corresponding plot is denoted as "Stratum specific influential plot". For these strata, the individual contributions to these quantities should be examined carefully to determine whether the cases and/or controls are the cause of the large values.

Standardized residuals *vs* leverage plot

Plotting the sum of squared Pearson residuals $S_k^* = \sum_j^{m_i} z_{kj}^2$ against respective leverage values to identify the outlying match sets (Moolgavkar *et al*., 1984) has been recommended and a methodology to identify the high leverage pairs and match sets with large standardized residuals was suggested. It was claimed that if there are d explanatory variables and n matched pairs in the model considered, matched pairs with leverage value greater than 2d/n are high leverage pairs and that matched pairs with standardized residuals greater than the number of controls in a stratum are worse fit by the model.

Residual analysis

Chi square probability plot of ordered residuals: It was suggested that the plot of ordered residuals (S_k) *vs* percentage points of a chi square distribution can be used to identify the poorly fit matched sets (Pregibon, 1984). Accordingly, linearity in the plots is the null configuration, whereas departures from linearity indicate outliers, and possibly the model differences suggesting that the current model should be augmented.

Lack-of-fit diagnostic plots: It was also suggested S_k^* as a diagnostic, where k denotes the k[th] strata and recommended plotting S_k^* and highlighting the matched sets where S_k^* greatly exceeds its estimated expected value of m_k (Moolgavkar *et al.,* 1984). In this illustration this plot will be denoted as "Stratum specific lack-of-fit statistic plot". Hosmer and Lemeshow(1989) defined the square of the standardized residual, r_{kj} as the lack-of-fit diagnostic (and denoted it as ΔX^2) for the j[th] observation in the k[th] stratum $\Delta x_{kj}^2 = r_{kj}^2$ and suggested plotting these values against their fitted values. Collett(2003) explained, that the fitted probabilities of the conditional likelihood model cannot be estimated since likelihood function does not involve the stratum parameters α_i's. So the fitted values obtained by equation 1 for the fitted β values are more like the contribution to the likelihood by the particular observation. Therefore in this illustration, instead of labeling the X axis as fitted values, it will be

labeled as the "Contribution to the likelihood". This plot is denoted as "Individual lack-of-fit statistic plot".

Computational statistics and programming: The user - friendliness of a computer programme is very important, since it helps the user to do the required analysis of diagnostic checking in this case, without wasting time. One approach is to facilitate the user by providing a graphical user interface (GUI). Therefore, Visual Basic for Applications (VBA) for Microsoft Excel 2003 is used since it permits a GUI for the programme (Hansen, 2004; Roman, 2002; Walkenbach, 2004). In addition Excel gives a massive grid view for the users (256 columns and 65536 rows). To get the parameter estimates and the design matrix for the calculation of goodness-of-fit statistics of a conditional logistics regression model, SAS 9.1 was used (Pennsylvannia State University, 2008; SAS Users Group International, 2008). A summary of the functionality of the programme is given in Figure 1.

The programme can be easily installed using Excel Add-ins manager. Data set should be opened in sheet 1 of an Excel worksheet before running the programme. The computer programme can be invoked using the Excel menu as in Figure 2 and the name of the programme is MATCHFIT. When the button MATCHFIT is clicked from the menu, the programme is started and, the startup window is given in Figure 3.

Depending on whether the model has only main effects or whether it has interactions the buttons labeled "main effects only model" or "main effects with interaction model" should be clicked respectively. If the user needs any help in this regard, a summary about the conditional logistic regression model with and without interaction terms can be found at the click of a button.

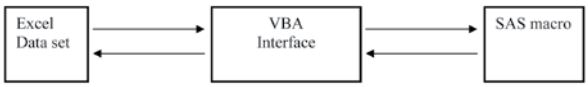


Figure 1: Summary of the functionality

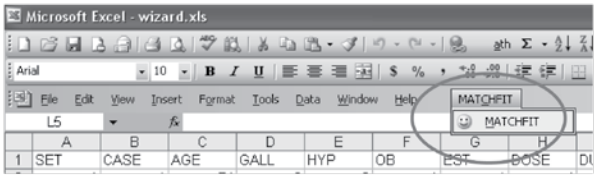


Figure 2: Starting MATCHFIT from the Excel menu bar

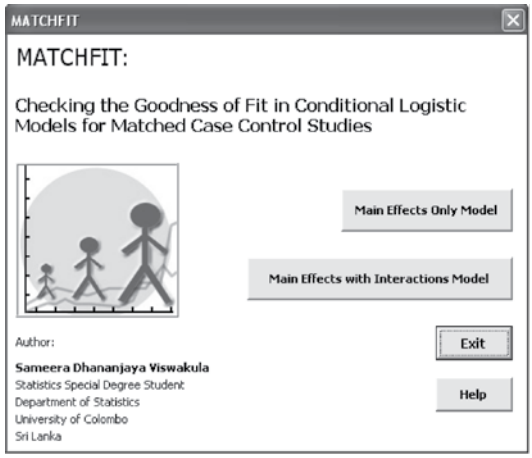


Figure 3: Startup window of the programme

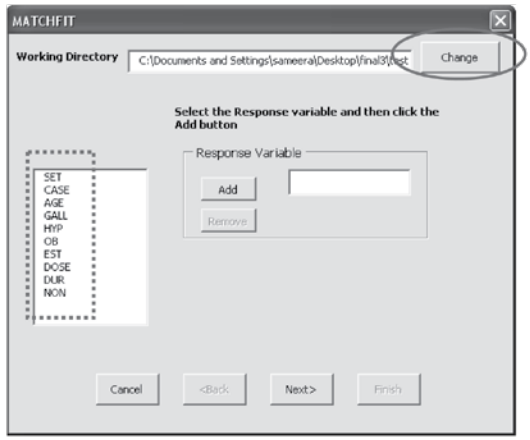


Figure 4: Main effects only model start up window

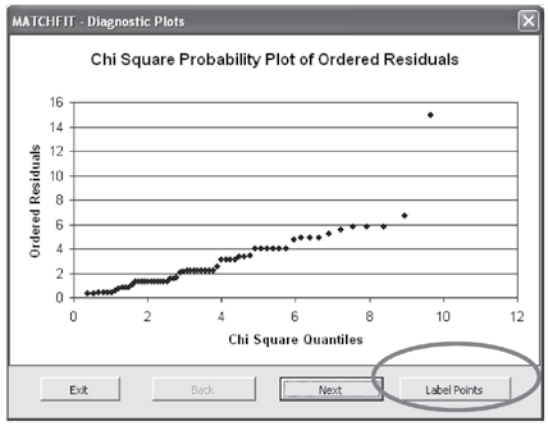


Figure 5: Diagnostic plots window

Table 1: Description of the variables in the data set

Variable Name	Description	Code/Range
SET	Matched set indicator	1 - 63
CASE	Case-control indicator	0 = Control 1 = Case
AGE	Age in years	55 -83
GALL	Gallbladder disease	0 = No 1 = Yes
HYP	Hypertension	0 = No 1 = Yes
OB	Obesity	0 = No 1 = Yes 9 = Unknown
EST	Estrogen usage	0 = No 1 = Yes
DOSE	Dose of conjugate	0 = 0 1 = 0.3 2 = 0.301-0.624 3 = 0.625 4 = 0.626-1.249 5 = 1.25 6 = 1.26-2.50 9 = Unknown
DUR	Duration of estrogen use (months)	0-95 96 = 96+ 99 = Unknown
NON	Non-estrogen drug	0 = No 1 = Yes

When the programme is started, it reads the Excel data set and gets user inputs to identify the response, stratum variable, continuous variables and categorical variables in the data set as in Figure 4.

Then the user has to specify the model which should be assessed. SAS® session is started from the VBA programme (Sastips, 2008). Using Object Linking and Embedding (OLE) automation (SAS Customer Support, 2008), variable information is passed to a SAS macro, to automate the process of fitting the model. All the variables are defined as global variables in the SAS macro [North East SAS Users Group (NESUG), 2008]. The user has to set the path of the tab delimited text file which has the relevant data set.

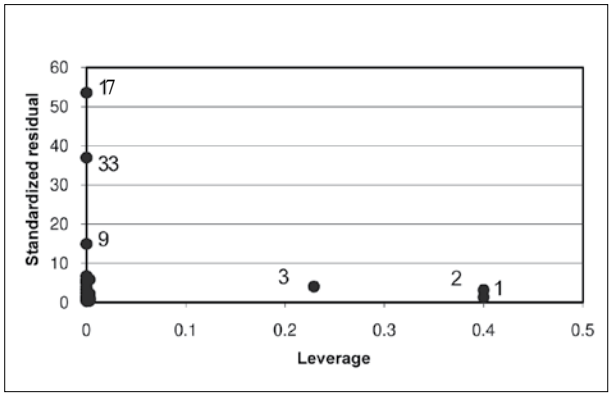
PROC LOGISTIC (SAS version 9.0) is used to fit the conditional logistic regression model with the STRATA statement. The technique used to get parameter estimates from the usual SAS output to a data step is output delivery system (ODS). The design matrix is taken to a data step using the relevant option. Then using dynamic data exchange (DDE) the parameter estimates and the design matrix is taken back to the Excel workbook. The parameter estimates are visible to the user through the VBA interface and if an error occurs during the SAS estimation process, the error can be traced from the SAS log file since ODS ON option is used in the SAS macro.

Then the relevant statistics are computed and the plots are automatically drawn in the VBA interface. If the user wants to identify the deviated observation in plots, the programme facilitates that requirement as well. This is illustrated in Figure 5 for the chi-square probability plot of ordered residuals. The SAS macro and the algorithms are available on request from the authors.

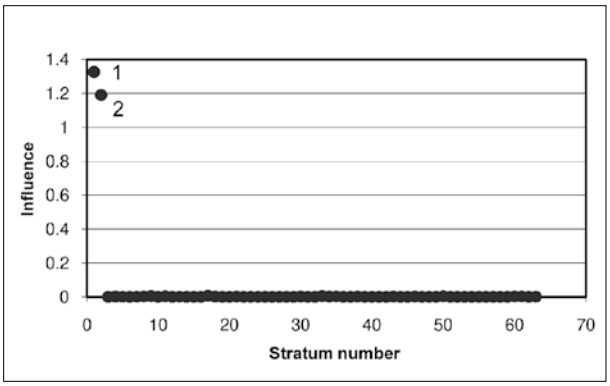
Illustration of programmes developed

Description of the data set

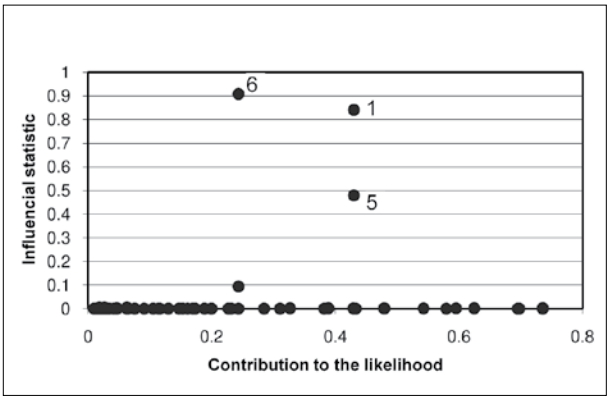
The methods described and implemented in previous sections are applied to a case control data set of leisure world study of endometrial cancer as related to treatment with estrogen for menopausal symptoms and other risk factors. This data set contains 315 records. One case is matched with four controls considering AGE as the confounding variable. Therefore, there are 63 matched sets and 10 variables in the data set. The leisure world study was done in Los Angeles (Mack *et al*.,1976). The variables in the data set are given in Table 1. Each matched pair consists of a case of endometrial cancer (outcome=1) and four controls (outcome=0) matched on the basis of age.



6(a): Standardized residuals *vs* leverage



6(b): Plot of stratum specific influence statistics



6(c): Plot of individual influence statistics

Figure 6: Plots for sensitivity analysis

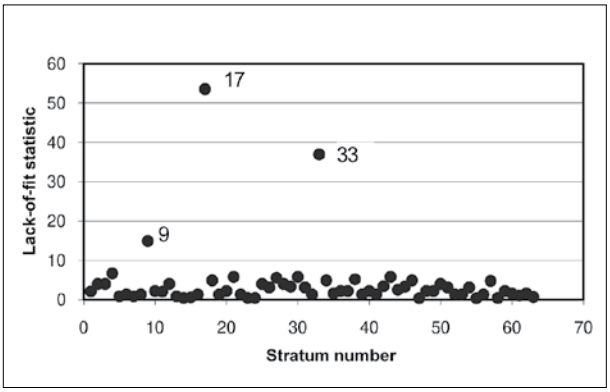
Results of data analysis

Here CASE is the response variable and AGE is the strata or the matching variable. Since the variable SET is the matched set indicator, it will not be used for modelling. Before going into advanced analysis, univariate analysis is done in order to identify the significant variables. As there are 8 explanatory variables and not all of these need to be significant and included in the model, it is important to have an intermediary stage where univariate analysis is used to identify whether the variables are significant at a liberal level [Collett, (2003) suggests a liberal level of 20%]. This is imperative if the number of explanatory variables is so large that the package does not allow fitting all the variables into the model due to memory constraints. In the univariate analysis, the associations between the response variable and each explanatory variable are considered, while other variables are not adjusted for. Chi square test of association is done, in order to find the significant variables at 20% significance level. Here a liberal significance level of 20% is considered because univariate analysis does not adjust for other variables.

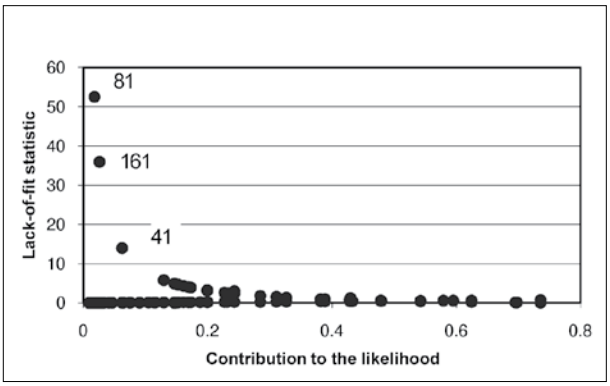
As all the explanatory variables are significant at the 20% level in the univariate analysis, every variable was considered in the modelling. The forward selection method (Agresti, 2002) is used for variable selection. When there is collinearity in the explanatory variables, none of the automatic selection methods including forward selection are appropriate for variable selection. Thus it is assumed that the collinearity in this data set is within manageable limits.
The final model selected is

$$\text{logit}\ \{P_j(X_{ij})\} = \alpha_j + \beta^{EST}_{ij} + \beta^{GALL}_{ij} \qquad ...(2)$$

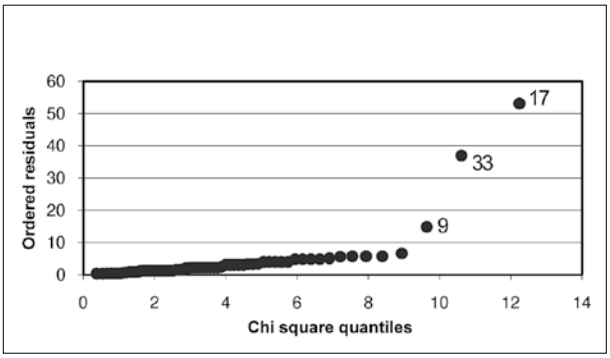
where i = 0, 1, 2,...,4 and j=1, 2, …63 and $P_j(X_{ij})$ is the probability of being a case in the j^{th} strata given explanatory variables X_{ij} where i corresponds to the i^{th} individual and j corresponds to the strata, α_j is the effect of the j^{th} strata, EST_{ij} is the estrogen usage indicator of the



7(a): Plot of stratum specific lack-of-fit statistics



7(b): Plot of individual lack-of-fit statistics



7(c): Chi square probability plot of ordered residuals

Figure 7: Plots for residual analysis

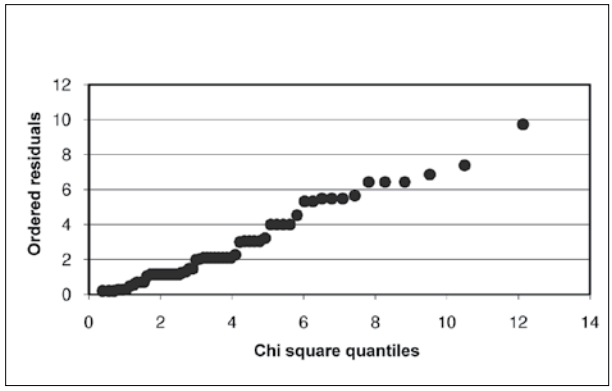


Figure 8: Chi Square probability plot of ordered residuals

i^{th} individual in the j^{th} strata and $GALL_{ij}$ is the gallbladder disease indicator for the i^{th} individual in the j^{th} strata.

Diagnostic checking

Goodness-of-fit of the model was checked using the computer programme, which has been developed.

Sensitivity analysis

Figure 6(a) plots standardized residuals versus leverage. For this data set, the value corresponding to 2d/n = 2*2/63 = 0.063492. Outlying observations corresponding to matched sets are labelled. Figure 6(a) indicates that matched sets 1, 2 and 3 have high leverage.

Observations with standardized residuals greater than the number of cases (4) in a stratum, are worse fit by the model. Therefore observations corresponding to matched sets 9, 17 and 33 result in large outliers [Figure 6(a)]. While observations corresponding to matched sets 3, 4, 12, 17,18, 21, 25, 27, 28, 30 and 34 correspond to lesser outliers as these are somewhat away from the remaining observations. Observations 1,2 and 3 have larger leverage values.

Figure 6(b) is a plot of the stratum specific influence statistic. Here the influence statistic is calculated for each case control matched set. According to previous studies (Moolgavkar *et al.,* 1985; Pregibon, 1984; Hosmer Lemeshow,1989) Figure 6(b) shows that matched sets 1 and 2 are highly influential to the model.

Figure 6(c) is a plot of individual influence statistic versus the contribution to the conditional logistic likelihood. Hosmer and Lemeshow(1989) claimed that observations with large values of these statistics would be judged to have large influence values. They proposed

that by carefully examining these large values it can be determined, whether case and/or controls are the cause for the large values. From Figure 6(c) it is evident that observations 1, 5 and 6 have significantly large influence values.

Table 2 shows the data information of the observation number 1, 5 and 6. The value for the response variable is 1 for the case and 0 for the controls. From Table 2 it is apparent that highly influential observation number 1 and 6 are cases whereas 5 is a control.

Residual analysis

Figure 7(a) is a plot of stratum specific lack-of-fit statistics. Matched case control sets are also known as strata and since the lack-of-fit statistic is calculated for each stratum, the term "Stratum-specific" is used. Therefore from Figure 7(a) it can be seen that, matched sets17 and 33 have significantly large lack-of-fit values. The matched set 9 also shows a considerable deviation from the model. This means the matched sets 9, 17 and 33 could be poorly fit by the model.

Table 2: Data information of the observation number 1, 5 and 6

OBS #	CASE	AGE	GALL	HYP	OB	EST	DOSE	DUR	NON
1	1	74	0	0	1	1	4	1	1
5	0	75	0	0	1	1	1	0	1
6	1	67	0	0	0	1	6	1	1

Table 3: Data description of the matched set number 17

SET	CASE	AGE	GALL	HYP	OB	EST	DOSE	DUR	NON
17	1	70	0	0	1	0	0	0	1
17	0	70	1	1	1	1	2	0	1
17	0	70	0	1	1	1	3	0	1
17	0	70	0	1	1	1	1	0	1
17	0	70	0	1	1	0	0	0	1

Table 4: Data information of the observation number 41, 81 and 161

OBS #	CASE	AGE	GALL	HYP	OB	EST	DOSE	DUR	NON
41	1	61	0	0	9	0	0	0	1
81	1	70	0	0	1	0	0	0	1
161	1	64	0	1	1	0	0	0	1

Table 5: Details of model 1 and model 2

Effect	**Statistic**	**Model 1**	**Model 2**
GALL	Estimate (level 0)	-0.653	-0.747
	Odds ratio	0.271	0.224
	Standard error	0.202	0.2172
	95% Confidence interval	(0.123 , 0.598)	(0.095, 0.526)
	p value	0.0012	0.0006
EST	Estimate (level 0)	-1.111	-1.5075
	Odds ratio	0.108	0.049
	Standard error	0.2299	0.2973
	95% Confidence interval	(0.044, 0.267)	(0.015, 0.157)
	p value	<0.0001	<0.0001

The data description of strata 17 is considered for further examination. Table 4 shows the data description of the 17th matched set.

In the 17th stratum three out of four controls were given estrogen, whereas the corresponding case in the stratum was not (Table 3). Excessive estrogen is associated with most of the risk factors that have been linked to endometrial carcinoma (Rose,1996). Therefore, controls in the stratum 17 have a high risk to be exposed to endometrial cancer than the case. This could be the reason for the lack-of-fit of the particular stratum.

The similar pattern can be observed for stratum 33 where all the controls are highly exposed to estrogen while the case is unexposed. This could be the reason for the poor fit of matched set number 33.

Figure 7(b) is a plot of individual lack-of-fit statistic versus the contribution to the conditional logistic likelihood by each observation. Hosmer and Lemeshow(1989) claimed that strata with large values of lack-of-fit statistic would be judged to be poorly fit. From Figure 7(b) it can be seen that observation 81, 161 and 41 are poorly fit by the model.Table 4 shows the data information of the observation number 41, 81 and 161. The value for the response variable is 1 for the case and 0 for the controls.

The individuals 41, 81 and 161 are cases who have not had gallbladder disease or given estrogen (Table 4). Therefore they may have developed endometrial cancer due to another risk factor (not because of the estrogen use or gallbladder disease). Figure 7(c) gives the chi square probability plot of ordered residuals.

A previous study (Pregibon, 1984) shows that linearity in the chi square probability plot indicates a well fitting model whereas a departure from linearity shows that the model has to be augmented. Figure 7(c) shows observations corresponding to matched sets 9, 33 and 17 deviate from the linearity. Since Figure 7(c) shows deviation of several strata from linearity, and according to Pregibon(1984) the current model should then be augmented, two-way interaction terms are added to the model [C] one at a time, where

$$\text{Model [C]} = \text{intercept} + \text{EST} + \text{GALL} \quad ...(3)$$

However, the results obtained indicated that none of the 2 factor interactions are significant at the 5% level. Agresti (2002) explains that the convention is that if the lower order interactions are non-significant then none of the higher order interactions are included in the model. Based on this argument as none of the two factor interactions were significant, higher order interactions were not tested. Therefore model C is considered as the best fitting model for this data.

From the standardized residuals versus leverage plot [Figure 6(a)], stratum specific lack-of-fit statistic plot [Figure 7(a)] and chi square plot of ordered residuals [Figure 7(c)], it is evident that observations 9, 17 and 33 are poorly fitted by the model.

Figures 6(a), 7(a) and 7(c) indicate that the matched sets 9, 17 and 33 are not well fitted by the model. Thus a model with these sets removed should be examined. Plots 6(b), 6(c) and 7(b) show individual statum specific statistics that are not well fitted by the model. If these individuals are removed from the data set then our matched case control study would no longer have a constant case to control ratio of 1: k. This situation cannot be handled by the methods developed in this paper, so these additional observations are retained in the new model.

Therefore the model is fitted with and without the outlying strata 9, 7, 33 in order to examine whether the conclusions arrived at are the same in both cases.

Parameter estimates

Table 5 gives the parameter estimates, odds ratios, standard errors, 95% confidence limits of the odds ratios and p values for the variables selected for model with all the data (Model 1) and for the model without strata 9, 17 and 33 (Model 2).

Model for the complete data set (Model 1)

Table 5 indicates that the effect of both GALL and EST are highly significant. Negative values for both parameter estimates indicate that when the values of the variables increase from 0 to 1 (absent to present) the odds of endometrial cancer increases.

The results show that the odds of having endometrial cancer for a person who does not use estrogen for menopausal symptoms versus a person who uses estrogen for menopausal symptoms is 0.108. Therefore, a woman who uses estrogen for menopausal symptoms has more than 9 times the odds of developing endometrial cancer than a woman who does not use estrogen for menopausal symptoms. It can be seen that 95 % confidence interval for the odds ratio is (0.044, 0.267). Since 1 is not included in the confidence interval the effect of estrogen on endometrial cancer is significant.

The odds of having endometrial cancer for a person who does not have gallbladder disease versus someone who has had gallbladder disease is 0.271. Therefore a person, who has had gallbladder disease, has nearly 4 times more odds of developing endometrial cancer than a person who does not have gallbladder disease for developing endometrial cancer. From the Table 5 it can be seen that, 95 % confidence interval for the odds ratio is (0.123, 0.598). Since 1 is not included in the confidence interval, this is also significant.

Model for the data without strata 9, 17, 33 (Model 2)

By removing these strata from the data set, the model is to be refitted and diagnostic plots are obtained.The strata 9, 17 and 33 are removed from the analysis and a conditional logistic regression model is refitted. Then the diagnostics are carried out from the computer programme that has been developed.

Since the chi square probability plot of ordered residuals is the major diagnostic plot for the fit of the model, the corresponding plot for the modified data set is shown in Figure 8. It is clear that the plot is linear and the model fits the data well.

Table 5 as expected indicates more highly significant results for model 2 than for model 1. This is because strata 9, 17 and 33 which were removed had cases without estrogen and controls with estrogen thus being in the opposite direction to the expected effect. Now the parameter estimates for both GALL and EST are higher in magnitude but have the same direction. Thus the conclusions reached from both models are the same indicating that both with and without the outlying strata the same conclusions are reached.

Therefore it can be concluded that, the best fitted model is:

$$\text{logit}\ \{P_j(X_{ij})\} = \alpha_j + \beta^{\text{EST}}{}_{ij} + \beta^{\text{GALL}}{}_{ij}$$

where i = 0,1,2,...,4 and j=1,2, …63, EST_{ij} is the indicator of estrogen use for the i^{th} individual in the j^{th} strata and $GALL_{ij}$ is the indicator of gallbladder disease for the i^{th} individual in the j^{th} strata.

RESULTS

From the analysis, it was found that, only the variables GALL and EST were significant in the model. None of the two factor interactions were significant. When the odds ratios were considered, it was found that, odds of getting endometrial cancer, for a person who has had gallbladder disease, is nearly four times than a person who does not have the gallbladder disease. The odds of having endometrial cancer, for a person who is exposed to estrogen is more than nine times that of a person who is not exposed to estrogen.

DISCUSSION AND CONCLUSION

SAS version 9.1 was chosen for the purpose of taking the parameter estimates and the design matrix of the conditional logistic regression model, which should be assessed. The main reason for choosing SAS 9.1 was that conditional logistic regression is available only for SAS 9.1 and newer versions. A user interface was designed in order to get the user inputs about the model and OLE automation was used to transfer the model and the data set information to the SAS system. A SAS macro was coded using global macro variables. These global macro variables were assigned to the PROC LOGISTIC procedure, so that SAS can fit the model for any number of variables. Then the parameter estimates and the design matrix of the considering model can be transferred to an Excel work book, using DDE method.

Using these Parameter estimates and the design matrix, goodness-of-fit statistics were calculated and the necessary diagnostic plots were drawn using VBA macro for Microsoft Excel. VBA for Excel 2003 is used as the programming language for this computer programme, the main reason for this choice being that VBA macros can be incorporated with Excel easily. Microsoft Excel 2003 gives a massive spread sheet support, so that users can store their data up to 256 columns and 65536 rows.

A conditional logistic regression model was fitted for the leisure world study data set, for identifying the factors affecting “endometrial cancer” as related to treatment with estrogen for menopausal symptoms and other risk factors.

Chi square association test and the Fisher’s exact test were used to identify the significant variables. All the variables (GALL, HYP, OB, EST, DOSE, DUR and NON) were significant at 20% level. Using the forward selection method, a conditional logistic regression model was fitted and only the variables GALL and EST were found to be significant at 5% level.

For testing goodness of the fitted model Proc Logistic is only capable of giving delta-beta plots which explain the influence of each observation on the parameters of the model. On the other hand plots obtained from the developed computer programme, in addition provide information on stratum specific lack-of-fit statistics.

Using the computer programme that has been developed, the goodness-of-fit of the model was assessed. From the chi square plot of ordered residuals, standardized residuals versus leverages plot and the stratum-specific lack-of-fit test, it was found that matched sets 9, 17 and 33 were significantly poorly fitted(outliers) by the model. When these sets were further analyzed, it was found that in these sets a majority of the controls were given estrogen whereas the case was not. Therefore these matched sets were removed and the model was refitted. In this case, the chi square probability plot of ordered residuals showed linearity. The conclusions arrived at by both models for the entire data set and the reduced data set were similar.

The plots used for goodness-of-fit testing of the model were very successful in identifying 3 outlying strata which were quite different from the other strata. In these 3 strata the case had not received estrogen whereas one or more control had received estrogen. This trend was contradictory to what is known about endometrial cancer in the literature.

Many popular statistical packages such as MINITAB and SPSS are incapable of doing conditional logistic regression. Two statistical packages that can do this analysis are SAS and EGRET of which the former is the more versatile. For testing goodness of the fitted model both packages are only capable of giving delta-beta plots which explain the influence of each observation on the parameters of the model. On the other hand plots obtained from the developed computer programme such as standardized residual versus leverage plot, plot of stratum specific influence statistics, plot of individual influence statistics, plot of stratum specific lack-of-fit statistics, plot of individual lack-of-fit statistics and chi-square probability plot of ordered residuals, in addition provide overall information on sensitivity and residual diagnostics.

The implemented software works only for the matched sets where there are equal numbers of controls matched with a single case in each stratum. This programme can be extended to cater to many-to-many matched case control studies (Kuruppumullage & Sooriyarachchi, 2007) and also for the models with higher order interaction terms.

References

1. Agresti A. (2002). *Categorical Data Analysis*. John Wiley & Sons Inc., New Jersey, USA.
2. Breshlow N.E. & Day N.E. (1980). *Statistical Methods in Cancer Research,* volume 1- The Analysis of Case-Control Studies. IARC Scientific Publications,Lyon, France.
3. Chapter 7 - Introduction to SAS Macros. *http://www.stat.psu.edu/~xzhan/stat597c/sp04/Chapter7.htm#top,* Accessed 16 May 2008.
4. Collett D. (2003). *Modelling Binary Data,* 2nd edition. Chapmann & Hall inc., London, UK.
5. Controlling SAS from Another Application Using OLE under Windows:Introduction to Automating SAS. SAS OnlineDoc 9.1.3 for the Web *http://support.sas.com/onlinedoc/913/getDoc/en/hostwin.hlp/introoleauto.htm.,* Accessed 2 June 2008.
6. Guidelines on Writing SAS® Macros for Public Use. *http://www2.sas.com/proceedings/sugi29/047-29.pdf*, Accessed 2 June 2008.
7. Hansen S.M. (2004). *Mastering Excel 2003 Programming with VBA*. Joel Fugazzotto, London, UK.
8. Hosmer D.W. & Lemeshow S. (1989). *Applied Logistic Regression*. John Wiley & Sons Inc., New York, USA.
9. Kuruppumullage D.P. & Sooriyarachchi M.R. (2007). Design and analysis technique for many-to-many matched case-control studies: an illustration using the Ille-et-Vilaine Database of esophageal cancer. *Sri Lankan Journal of Applied Statistics* **8**: 55-70.
10. Lazy Programmer Case Study: Dynamic Macro Code to Deal with Changing Number of Variables Over Time. *http://www.lexjansen.com/cgi-bin/xsl_transform.asp?x=sad&s=nesug_s&c=nesug*, Accessed 2 June 2008.
11. Mack T.H., Pike M.C., Henderson B.E., Pfeffer R.I., Gerkins V.R., Arthur B.S. & Brown S.E. (1976). Estrogen and endometrial cancer in a retirement community. *New England Journal of Medicine* **294**(23): 1262-1267.
12. Moolgavkar S.H., Lustbader E.D. & Venzon D.J. (1984). A geometric approach to nonlinear regression diagnostics with application to matched case-control studies. *Annals of Statistics* **12**(3): 816-826.
13. Moolgavkar S.H., Lustbader E.D. & Venzon D.J. (1985). Assessing the adequacy of the logistic regression model for matched case-control studies. *Statistics in Medicine* **4**(4): 425-435.
14. Pregibon D. (1984). Data analytic methods for matched case-control studies. *Biometrics* **40**(3): 639-651.
15. Roman S. (2002). *Writing Excel Macros with VBA*, 2nd edition, p.576. O'Reilly Media, California, USA.
16. Rose P.G. (1996). Endometrial Carcinoma. *The new England Journal of Medicine* **335**: 640-649.
17. Run SAS code from Excel with VB. *http://www.sastips.com/out.php?url=http%3A%2F%2Fwww.woodstreet.org*.uk, Accessed 10 April 2008.
18. Schlesselman J. (1982). *Case-Control Studies Design, Conduct, Analysis*. Oxford University Press, New York, USA.
19. Walkenbach J. (2004). *Excel 2003 Power Programming with VBA*. John Wiley & Sons Inc., New Jersy, USA.