JOURNAL OF APPLIED QUANTITATIVE METHODS

Applied Quantitative Methods in Medicine

STATISTICAL MODELING OF THE INCIDENCE OF BREAST CANCER IN NWFP, PAKISTAN

Salah UDDIN
PhD, University Professor, Chairman, Department of Statistics,
University of Peshawar, Peshawar, NWFP, Pakistan

E-mail: salahuddin_90@yahoo.com

Arif ULLAH
Lecturer in Statistics, Higher Education Department, Peshawar, NWFP, Pakistan

E-mail:

NAJMA
Lecturer in Statistics, Frontier Women University, Peshawar, NWFP, Pakistan

E-mail:

Muhammad IQBAL
Lecturer, Department of Statistics,
University of Peshawar, Peshawar, NWFP, Pakistan

E-mail:

Abstract: *Breast cancer is the most common form of cancer that affects women. It is a life threatening disease and the most common malignancy in women through out the world. In this study an effort has been made to determine the most likely risk factors of breast cancer and to select a parsimonious model of the incidence of breast cancer in women patients of the age 50 years and above in the population of North West Frontier Province (NWFP), Pakistan. The data were collected from a total of 331 women patients, arriving at Institute of Radiotherapy and Nuclear Medicine Peshawar, NWFP, Pakistan.*

*Logistic regression model was estimated, for breast cancer patients, through backward elimination procedure. Brown tests were applied to provide an initial model for backward elimination procedure. The logistic regression model, selected through backward elimination procedure contains the factors Menopausal status (M), Reproductive status (R), and the joint effect of Diet and family History (D*H). We conclude that menopausal status; reproductive status and the joint effect of diet and family history were the important risk factors for the breast cancer.*

*Separate models were then fitted for married and unmarried breast cancer patients. The best-selected model for married females is of factors Feeding (F), R, M, (D*H), whereas the best selected model for unmarried females has only one main factor Menopausal status. We conclude that breast feeding, reproductive status, menopausal status and the joint effect of diet and family history were the important risk factors of breast cancer in married women and the menopausal status was the important risk factor of breast cancer in unmarried women.*

Key words: *Logistic regression; backward elimination procedure; Brown method; Wald statistic*

1. Introduction

Cancer of breast is a disease that instills feelings of dread and fear in many women. Not only is it a life threatening disease, but it affects a part of the body that is central to women's sense of womanliness and femininity. It is a complex disease with the causes not yet fully understood. It is most likely caused by a number of factors interacting with each other, rather than by any one factor. The main identified risk factor of breast cancer is age, with woman aged 80 years or over being most at risk (Jelfs, 1999).

According to Australian Institute of Health and Welfare report (AIHW, 1998), 43% of breast cancer cases were in women between the ages of 45 and 64, and 22% were in women between the ages of 65 and 74 during 1982-1994. Approximately 18% of cases occurred in women younger than 45 years and women older than 74 years.

Before the age of 75, one in eleven women in Australia is expected to diagnose with breast cancer. In 1996, 9556 new cases of breast cancer were diagnosed in Australia and there were 2,623 deaths attributed to breast cancer. Similarly, of the 30201 deaths from breast cancer in Australian women from 1982-1994, 38% occurred in women aged 45 to 64, 28% in women aged 75 or over, 24% in women aged 65 to 74, and 10% in women younger than 45 years of age (Kricker and Jelfs, 1996).

For the period 1996-2000, women aged 20-24 have the lowest incidence rate, 1.4 cases per 100,000 population; women aged 75-79 have the highest incidence rate, 499 cases per 100,000 (Ries et al., 2003). Breast cancer incidence rates among African American women range from 89.8 in Rhode Island to 147.6 in Alaska (Hotes et al., 2003).

Until now no such statistical study has been made in the province of NWFP on the various risk factors of breast cancer. In this study an effort has been made to model the relationship between breast cancer in the population of NWFP and its probable risk factors.

2. Data Set

The data for this study were collected from Institute of Radiotherapy and Nuclear Medicine (IRNUM), Peshawar. The study is based on a sample of size 331 women, including 123 cases (breast cancer patients) and 208 control (not breast cancer patients) groups. Out of 331 women breast cancer patients, 31 (9.37%) patients are unmarried and 300 (90.63%) patients are married.

The suggested risk factors for fitting the model are family history (H), reproductive status (R), breast-feeding (F), oral contraceptives (C), menopausal status (M) and diet (D). The response variable for the study is the diagnosis of patient with breast cancer or not.

3. Method and Materials

Generalized linear models introduced by Nelder and Weldderburn (1972) are a class of statistical models, which is the natural generalization of classical linear model. It includes response variables that follow any probability distribution in the exponential family of distributions. An excellent treatment of generalized linear models is presented in Agresti (1996). In this study the response variable is binary; therefore, the logistic regression model is an appropriate model, which is a part of generalized linear models.

The response variable in logistic regression is usually dichotomous, that is, the response variable can take the value 1 with a probability of success θ, or the value 0 with probability of failure 1-θ. This type of variable is called a Bernoulli (or binary) variable.

The relationship between the predictor and response variables is not a linear function in logistic regression; instead, the logistic regression function is used, which is given as

$$\theta(x) = \frac{e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}}{1 + e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... \beta_k x_k)}} \quad (1)$$

We now find the link function for which the logistic regression model is a generalized linear model (GLM). For this model the odds of making response 1 are

$$\frac{\theta(x)}{1-\theta(x)} = e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)} \quad (2)$$

$$\log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \quad (3)$$

Thus the appropriate link is the log odds transformation, the logit. The logistic regression model is given by

$$\text{logit}\,[\theta(x)] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \quad (4)$$

The parameters in this model, α, β_1, β_2... β_k can no longer be estimated by least squares, but are found using the maximum likelihood method (Collett, 1991; Cox & Snell, 1989).

Logistic regression calculates the probability of success over the probability of failure; therefore, the results of the analysis are in the form of an odds ratio. Logistic regression also provides knowledge of the relationships and strengths among the variables.

The Wald statistic is commonly used to test the significance of individual logistic regression coefficients for each independent variable. The Wald statistic for the β_j coefficient is:

$$\text{Wald} = \left[\frac{\hat{\beta}_j}{S.E(\hat{\beta}_j)}\right]^2,$$

It is distributed as chi-square with 1 degree of freedom. The Wald statistic is simply the square of the (asymptotic) *t*-statistic. The Wald statistic can be used to calculate a confidence interval for β_j. We can assert with $100(1-\alpha)\%$ confidence that the true parameter lies in the interval with boundaries $\hat{\beta} \pm Z_{\alpha/2}(ASE)$, where ASE is the asymptotic standard error of logistic $\hat{\beta}$. Parameter estimates are obtained using the principle of maximum likelihood; therefore hypothesis tests are based on comparisons of likelihoods or the deviances of nested models. The likelihood ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1) \quad (5)$$

This log transformation of the likelihood functions yields a chi-squared statistic. This is the recommended test statistic to use when building a model through backward elimination procedure. Once $\hat{\beta}$ has been obtained, the estimated value of the linear systematic component (also known as linear predictor) of the model is

$$\hat{\eta}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} \quad (6)$$

From equation (6), the fitted probabilities $\hat{\theta}_i$ can be found using

$$\hat{\theta}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}.$$

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. Several different options are available during model creation. Variables can be entered into the model in the order specified by the researcher or logistic regression can test the fit of the model after each coefficient is added or deleted (backward elimination procedure), called stepwise procedure. Backward elimination procedure appears to be the preferred method of exploratory analyses, where the analysis begins with a full or saturated model and variables are eliminated from the model in an iterative process. The fit of the model is tested after the elimination of each variable to ensure that the model still adequately fits the data. When no more variables can be eliminated from the model, the analysis has been completed.

4. Analyses and Interpretation

We begin with the initial model having factors F, M, R, (F*D), (M*C), (D*C), (D*H), and (F*D*R), provided by Brown test. Using backward elimination method through SPSS package, the final model was selected at step 6 which contains two main factors (M and R) and one interaction factor (D*H). Thus the significant factors are M, R, (D*H). It means that menopausal status (M), reproductive status (R) and the joint effect of diet and family history (D*H) were the important risk factors for the breast cancer.

Table 1. Variables in Model 1

Variables	$\hat{\beta}$	S.E($\hat{\beta}$)	Wald	d.f.	P-value	Exp($\hat{\beta}$)	95% C.I. for Exp($\hat{\beta}$)	
							Lower	Upper
(D*H)	2.235	0.820	7.432	1	0.006	9.350	1.874	46.644
M	2.449	0.472	26.907	1	0.000	11.576	4.589	29.203
R	1.298	0.328	15.608	1	0.000	3.660	1.923	6.967
Constant	-1.083	0.147	54.361	1	0.000	0.339	-	-

The fitted model is:

$$\text{Logit}(\hat{p}) = -1.083 + 2.449M + 1.298R + 2.235(D*H) \quad (7)$$

4.1 Analysis According to Marital Status

Some factors like reproductive status (R), breast-feeding (F) and oral contraceptives (C) are irrelevant to unmarried women. Therefore, Separate models are fitted for married and unmarried women patients.

(a) Model for Married Women

The suggested risk factors (explanatory variables) in this case are family history (H), reproductive status (R), breast-feeding (F), oral contraceptives (C), menopausal status (M) and diet (D). We repeat the same process, starting from Brown method and then backward elimination procedure. The final model selected at step 5, contains the factors F, M and (D*H). It means that one new factor breast-feeding (F) is turned out significant in this case and other significant factors M, (D*H) are the same.

Table 2. Variables in Model 2

Variable	$\hat{\beta}$	S.E($\hat{\beta}$)	Wald	d.f	P-value	Exp($\hat{\beta}$)	95% C.I. for Exp ($\hat{\beta}$)	
							Lower	Upper
(D*H)	1.746	0.843	4.001	1	0.045	5.730	1.036	31.701
M	2.382	0.525	20.574	1	0.000	10.823	3.867	30.291
F	1.342	0.368	13.287	1	0.000	3.827	1.860	7.875
Constant	-1.129	0.170	44.300	1	0.000	0.323	-	-

The fitted model is:

$$\text{Logit}(\hat{p}) = -1.129 + 1.342F + 2.382M + 1.746\,(D*H) \quad (8)$$

(b) Model for Unmarried Women

The suggested risk factors in this case are diet (D), menopausal status (M) and family history (H). We begin with the initial model having factors D, M & H, provided by Brown method for backward elimination method. The procedure selects the final model at step 1 with only one main factor M. Hence the menopausal status (M) is the important risk factor of breast cancer in unmarried case.

Table 3. Variables in Model 3

Variable	$\hat{\beta}$	S.E($\hat{\beta}$)	Wald	d.f	P-value	Exp($\hat{\beta}$)	95% C.I. for Exp ($\hat{\beta}$)	
							Lower	Upper
M	2.748	1.090	6.354	1	0.012	15.610	1.843	132.222
Constant	-0.671	0.256	6.856	1	0.009	0.511	-	-

The fitted model is

$$\text{Logit}(\hat{p}) = -0.671 + 2.748\,(M) \quad (9)$$

In this model the variable 'M' is selected as an important factor. This model is the reduce form of model 1 and model 2.

5. Conclusion

The purpose of this study was to estimate a model to determine the most likely risk factors of breast cancer, women patients of age 50 years or above, in NWFP. The phenomena of breast cancer was studied in relation to different risk factors namely, oral contraceptives(C), diet (D), menopausal status (M), family history (H), breast feeding (F) and reproductive status (R) of 331 patients arriving at IRNUM Peshawar. Out of these 331 patients, the number of breast cancer patients (case group) were 123 and 208 had no breast cancer (control group); 31 (9.37%) were unmarried and 300 (90.63%) were married. Older women were at increasing risk over time.

For model fitting procedure we used a binary response variable B (Breast cancer), taking the value 1 for breast cancer patients and 0 otherwise. Brown method was used for selecting initial model. Backward elimination procedure was used to determine a parsimonious model. The logistic regression analysis was then applied to the data.

The variables chosen initially as predictors were R, H, C, D, M and F. The logistic regression model, selected through backward elimination procedure contains the factors M, R and the interaction term (D*H). It means that menopausal status; reproductive status and the joint factor (D*H) were the important risk factors for the breast cancer. Separate logistic regression model was then fitted for married and unmarried women patients. For married females, we obtained the model with predictors F, M, (D*H). Thus breast feeding, menopausal status and (D*H) were the important risk factors. For unmarried women on the other hand, we obtained the final model containing only one main factor M. Hence the menopausal status was the only important risk factor.

Finally on the basis of analysis based on sample of 331 patients, we concluded that menopausal status, reproductive status and joint effect of diet and family history are the important risk factors. While in addition to these factors, breast-feeding is also the most important factor in the case of married patients. Therefore, one of our main findings is that breast-feeding is a preventive measure against breast cancer. Furthermore, the risk of breast cancer is found to be increasing with age. The highest incidents among women were in the age group (40-59). Therefore, it is suggested that breast cancer screening may be advised before this age group.

References

1. Agresti, A. **An Introduction to Categorical Data Analysis**, John Wiley and Sons Inc., New York, 1996
2. AIHW **Breast Cancer Survival in Australian Women 1982-1994**, Canberra, Australian Institute of Health and Welfare, 1998, online source: www.aihw.gov.au/publications/health/bcsaw82-94/bcsaw82-94.pdf
3. Collett, D. **Modeling Binary Data**, Chapman & Hall, London, 1991
4. Cox, D. R. and Snell, E. J. **Analysis of Binary Data**, 2nd edition, Chapman and Hall, London, 1989
5. Hotes, J. L., Mclaughlin, C. C., Lake, A., Frith, R., Roney, D., Cormier, M., Eulton, J. P., Holowaty, E., Howe, H. L., Kosary, C. and Chen, R. W. (eds.), **Cancer in North American 1996–2000, Volume No. 1: Incidence, Volume No. 2: Mortality**, Springfield IL, and North American Association of Central Cancer Registries, 2003
6. Jelfs, P. **Breast caner in Australia: an overview, The Breast Cancer Bulletin 1999**, 1999, pp. 1-2

7. Kricker, A. and Jelfs, P. **Breast Cancer in Australian Women 1921-1994**, Canberra, Australian Institute of Health and Welfare, 1996, pp. 12
8. Nelder, J. A. and Wedderburn, R. W. M. **Generalized Linear Models**, Journal of Royal Statistical Society, Series A, 135, 1972, pp. 370-384
9. Ries, L. A. G., Eisner, M.P. and Kosary, C. L. (eds.) **SEER Cancer Statistics Review, 1975-2000**, Bethesda, MD: National Cancer Institute, 2003