RESEARCH COMMUNICATION

Association Between Risk of Breast Cancer and Fertility Factors - a Latent Variable Approach

Mohamad Amin Pourhoseingholi[1], Yadolah Mehrabi[1], Hamid Alavi-Majd[2], Parvin Yavari[3], Azadeh Safaee[1]*

Abstract

Background: Breast cancer is the most common malignant tumor in females. Many studies have been carried out in order to assess the reproductive risk factors. Particular attention has focused on information regarding fertility, including breastfeeding, age at first birth and number of live births. These factors are highly correlated with each other. The objective of this study was to employ latent variables to reduce the confounding effect of this correlation with a logistic regression analysis. Methods: The investigation drew upon results from a dataset belonged to a hospital based case-control study covering 303 breast cancer patients and 303 hospital controls. Data were collected through interview and reproductive variables included age at first full-term pregnancy and live birth, number of pregnancies and live births, and total length of breast feeding. Latent variables were generated using factor analysis and principal components analysis. Results: The study revealed that for both latent variable approaches the odds ratios of two latent variables significantly indicated a protective impact of number of pregnancy and live birth and breastfeeding and a prognostic relation with age at first pregnancy or live birth. Conclusion: The findings suggest that breastfeeding and decreasing age at first live birth have protective influences on breast cancer risk. Also using statistical model with latent variables in the presence of collinear data leads to reliable results.

Key Words: Breast cancer - fertility factors - latent variables

Asian Pacific J Cancer Prev, **9**, 309-312

Introduction

Breast cancer is the most common malignant tumor in females, affecting one in nine healthy women (Tarone, 2006) the incidence rising rapidly from the 1970s to the 1990s in most countries (Althius et al., 2005). Informal data indicated that breast cancer has increased in Iran and since 1999 its incidence has ranked highest in the Iranian cancer registry data (Shamsa et al., 2002).

Many studies have been cried out in order to asses the reproductive risk factors of breast cancer. One of the most considerable associated factors is the prior information regarding to fertility consist of breastfeeding, age at first birth and number of live birth. In addition these factors are the few protective ones that potentially modifiable risk factors for breast cancer (Collaborative Group., 2002, Ursin et al., 2006, Berkowitz et al., 1990, Althuis et al., 2004).

Logistic regression is a model used for prediction of the probability of occurrence of an event. It makes use of several predictor variables that may be either numerical or categories. The best regression models are those in which the predictor variables each correlate highly with the dependent (outcome) variable but correlate at most only minimally with each other. But when there are many explanatory variables of interest, the efficiency of the model is reduced, especially if there are also strong relationships among independent variables i.e. multicollinearity (Chattergee et al., 2002). Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn mean that coefficients for some independent variables may be found not to be significantly different, whereas without multicollinearity and with lower standard errors, these same coefficients might have been found to be significant and the researcher may not have come to null findings.

*[1]Department of Health System Research, Research Center for Gastroenterology and Liver Disease, [2]Department of Biostatistics, [3]Department of Health and Community Medicine, Shaheed Beheshti Medical University, Iran *For Correspondence: 7th floor of Taleghani Hospital, Tabnak Street, Evin, Tehran, Iran. Email: azadesafaee@yahoo.com, Amin_phg@yahoo.com*

One solution in this case is removing the most inter correlated variable(s) from analysis. This method is misguided if the variables were there due to the theory of the model.

In some cases, variables involved in multicollinearity can be combined into a single variable called a latent variable. Latent variables are such variables that not be observed directly but can be generated by a transformation of other observed variables and employed instead of original collinear explanatory variables (Van Eye et al., 1994). The use of latent variables for data reduction actually has had more application in psychology and social sciences (Kenneth., 2002) but because of some conditions in the areas of epidemiology and medical sciences, in which researchers often encounter situations where there are many variables related to each other, the problem of multicollinearity is expected to occur (Kleinbaum., 1994). The latent variable technique can then be employed to overcome the problem. However, multicollinearity makes problems in logistic regression too, so researchers just focus on linear regression with normal responses. The objective of the present study was to employ latent variables to reduce the effect of multicollinearity in the analysis of reproductive fertility factors for breast cancer.

Materials and Methods

The present investigation draws upon results from a dataset belonging to a hospital based case-control study that was carried out to elucidate factors associated with development of breast cancer (Yavari et al., 2005). The total sample comprised 303 breast cancer patients and 303 hospital controls. All the cases and controls were selected from a teaching university hospital in North Tehran. Data were collected through interview using structured questionnaires and reproductive variables were included age at first full-term pregnancy (AFP) and the age at first live birth (AFLB), number of pregnancies (NP) and number of live births (NLB), and the total length of breastfeeding (TLBF). The term VIF (variance inflation factor) and the Pearson's correlation coeeficient were employed to indicate the measure of collinearity. Then a conditional logistic regression was performed for analysis, first using all original fertility variables and second using latent variables. Latent variables were generated using factor analysis and principal components analysis.

Table 1. The Correlation Matrix and VIF of the Five Fertility Variables

(VIF)*	AFLB	AFP	TLBF	NLB	NP	
4.57	-0.42**	-0.44**	0.67**	0.90**	1	NP
5.83	-0.47**	-0.47**	0.75**	1		NLB
2.13	-0.41**	-0.42**	1			TLBF
15.41	0.97**	1				AFP
15.34	1					AFLB

*Variance Inflation Factors **Significant at the 1 % level

Table 2. Estimated OR and 95% CIs of Model Using Logistic Regression with Original and Latent Variables

Models	Odds Ratio	CI for Odds Ratio	P-value
Logistic Regression with Original Variables			
Intercept	1.1		0.62
NP	1.06	(0.89 - 1.27)	0.50
NLB	0.84	(0.50-1.40)	0.51
TLBF	0.80	(0.59-1.08)	0.14
AFP	0.000029		<0.001
AFLB	67960	(10184-453503)	<0.001
Logistic Regression with Factor Scores			
Intercept	0.96		0.64
Factor 1	0.76	(0.64-0.91)	0.002
Factor 2	1.33	(1.12-1.59)	0.001
Logistic Regression with Principal components			
Intercept	0.96		0.64
Component 1	0.77	(0.65-0.91)	0.002
Component 2	1.31	(1.10-1.56)	0.003

Results

Table 1 presents the high linear correlation among the fertility factors investigated, with alarge VIF. VIF > 5 indicates high multicollinearity for all factors which are dramatically high for NLB, AFP and AFLB and moderate for NA and TLBF.

First an ordinary multiple logistic regression was applied to these variables without attention to this clear problem. From Table 2, it is found that when all variables included in the model only the estimate for AFP and AFLB are statistically significant, and others are not with very large p-values. But the results for these two significant variables show an unusual odds ratio (approximately near zero for AFLB and OR=67960 for AFP). These unexpected results caused by the high correlation among the variables, making interpretation too difficult and model is not reliable at all. In order to ignore the effect of multicollinearity we conducted two latent variables approach; factor analysis and principal components analysis. The correlation structure and what we have known substantively about the observed variables suggest creating two latent variables.

The value for the coefficient derived from both methods (not shown here) indicated that the first latent variable (Factor1 and Component1 in both techniques) has been weighted with NP, NLB and TLBF but the second one is a linear combination of AFP and AFLB. This is due to the fact that all analysis in these techniques are based on the structure of correlation existing among original dataset and the results appear as a component where the greatest variance by any projection of the data comes to lie on it or as a linear combination of the observed data, plus error terms, called factor. Therefore the first latent variable is an index of number of pregnancy, live birth and breastfeeding respectively and the second latent variable is an index of age at first pregnancy or live birth.

Then a logistic regression was carried out with these two latent variables. From Table 2, we see that for both latent variable approaches the odds ratios of two latent variables are significant indicated a protective impact of first latent variable (OR=0.76, 95% CI=0.64-0.91 for factor analysis and OR=0.77, 95% CI=0.65-0.91 for principal component analysis) and a prognostic risk factor for second one (OR=1.33, 95% CI= 1.12-1.56 for factor

analysis and OR=1.31, 95% CI=1.10-1.56 for principal component analysis).

Discussion

The aim of this study was employed latent variable approach to handle the collinearity problem in order to analyze the relationship between fertility factors and breast cancer. Our findings indicated that although it seems that the fertility factors didn't affect breast cancer in a logistic model included all variables, the alternative model with latent variables showed significant protective impact for first latent variable as an index of number of pregnancy, live birth and breastfeeding, and a significant prognostic results for second latent variable as an index of age at first pregnancy or live birth.

Breastfeeding was strongly inversely associated with breast cancer risk (Sumitra Shantakumar et al., 2007) reduced the risk of breast cancer (Ma et al., 2006) and Lack of breastfeeding, is significantly associated with breast cancer (Faheem et al., 2007).

Mahouri et al reported a significant relation with breastfeeding in a case control study in Iran (Mahouri et al., 2007). A literature review by Lipworth et al concluded that women who nursed their children had a reduced incidence of breast cancer compared to those who did not, with a clear inverse relationship between duration of lactation and breast cancer incidence (Lipworth et al., 2000). Additionally, there has been a trend toward delaying pregnancy until later in life (Berkowitz et al., 1990) and thus the issue of subsequent pregnancy and breastfeeding is becoming more relevant and worth consideration. Some recent study depicted that Women in the oldest age at first birth category were on average at a 27% greater risk of breast cancer (Ma et al., 2006).

Age at first live birth was another associated factor, reduces the risk of breast cancer (Titus-Ernstoff et al., 2006). As this age decreases the risk of breast cancer decreases respectively. Number of live birth was another prognostic factor that affects the risk of breast cancer. The protective effects of a greater number of births and an early age at first birth against breast cancer suggest that their effects influence risk predominantly through hormonal mechanisms that involve estrogen and progesterone. The effects of these hormones on breast tissue depend upon the amount of both hormones and their specific receptors (Anderson., 2002, Dickson et al., 2000). However some study finds no significant relation with number of live birth and breast cancer (Kishk., 1999), this item is in direct relation to number of children breastfed and the risk decreased as the number of children breastfed increased (Enger et al., 1998) so the impact of this item appeared clearly in the first latent factor.

In this paper we concluded that using latent variable approach leads to better interpretation and conclusion than ordinary logistic regression in the presence of multicollinearity. We also showed in a simulation study that this approach is more power full and reliable than models with original collinear data (Pourhoseingholi et al., 2008). Although latent variable technique has more application in psychology, in the field of medical sciences researchers should be interested in this technique according to its advantage and applications.

References

Althuis MD, Dozier JM, Anderson WF, Devesa SS, Brinton LA (2005). Global trends in breast cancer incidence and mortality 1973-1999. *IEA*, **34**, 405-12.

Althuis MD, Fergenbaum JH, Garcia-Closas M, Brinton LA, Madigan MP, Sherman ME (2004). Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. *Cancer Epidemiol Biomarkers Prev*, **13**, 1558-68.

Anderson E (2002). The role of oestrogen and progesterone receptors in human mammary development and tumorigenesis. *Breast Cancer Res*, **4**, 197-201.

Berkowitz GS, Skovron ML, Lapinski RH, Berkowitz RL (1990). Delayed child-bearing and the outcome of pregnancy. *N Engl J Med*, **322**, 639-64.

Chattergee S, Hadi AS, Price B (2002). Regression Analysis by Example. John Wiley & Sons, USA. pp: 225-8.

Collaborative Group on Hormonal Factors in Breast Cancer (2002). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50,302 women with breast cancer and 96,973 women without the disease. *Lancet*, **360**, 187-95.

Dickson RB, Stancel GM (2000). Estrogen receptor-mediated processes in normal and cancer cells. *J Natl Cancer Inst Monogr*, **27**, 135-45.

Enger SM, Ross RK, Paganini-Hill A, Bernstein L (1998). Breastfeeding experience and breast cancer risk among postmenopausal women. *Cancer Epidemiol Biomarkers Prev*, **7**, 365-9.

Faheem M, Khurram M, Jafri IA, et al (2007). Risk factors for breast cancer in patients treated at NORI Hospital, Islamabad. *J Pak Med Assoc*, **57**, 242-5.

Kenneth AB (2002). Latent variables in psychology and the social sciences. *Annu Rev Psychol*, **53**, 605-34.

Kishk NA (1999). Breast cancer in relation to some reproductive factors. *J Egypt Public Health Assoc*, **74**, 547-66.

Kleinbaum D (1994). Logistic Regression.Springer, New York. pp: 168.

Lipworth L, Bailey LR, Trichopoulos D (2000). History of breastfeeding in relation to breast cancer risk: a review of the epidemiological literature. *J Natl Cancer Inst*, **92**, 302-12.

Ma H, Bernstein L, Pike MC, Ursin G (2006). Reproductive factors and breast cancer risk according to joint estrogen and progesterone receptor status: a meta-analysis of epidemiological studies. *Breast Cancer Res*, **8**, R43.

Mahouri K, Dehghani Zahedani M, Zare S (2007). Breast cancer risk factors in south of Islamic Republic of Iran: a case-control study. *East Mediterr Health J*, **13**, 1265-73.

Pourhoseingholi MA, Mehrabi Y, Alavi-Majd H, Yavari P (2008). Using latent variables in logistic regression to reduce multicollinearity, A case-control example: Breast cancer risk factors. Ital J PH; VI, n 1, Spring, under published.

Shamsa AZ, Mohagheghi MA (2002). Final report of the National project for cancer registry. Islamic Republic of Iran.

Shantakumar S, Terry MB, Teitelbaum SL, Britton JA, Millikan RC, Moorman PG et al (2007). Reproductive factors and breast cancer risk among older women. *Breast Cancer Res Treat*, **102**, 365-74.

Tarone RE (2006). Breast cancer trends among young women in the United States. *Epidemiology*, **17**, 588–590.

Titus-Ernstoff L, Tosteson AN, Kasales C, et al (2006). Breast

cancer risk factors in relation to breast density (United States). *Cancer Causes Control*, **17**, 1281-90.
Ursin G, Sun CL, Koh WP, et al (2006). Associations between soy, diet, reproductive factors, and mammographic density in Singapore Chinese women. *Nutr Cancer*, **56**, 128-35.
Van Eye A, Clogg CC (1994). Latent variables analysis; application for developing research; SAGE publication. PP: 3-35.