Using latent variables in logistic regression to reduce multicollinearity, A case-control example: breast cancer risk factors

Mohamad Amin Pourhoseingholi[1], Yadolah Mehrabi[2], Hamid Alavi-Majd[3], Parvin Yavari[4]

[1]Research Center for Gastroenterology and Liver disease Shaheed Beheshti Medical University, Iran; [2]Shaheed Beheshti Medical University, Department of Biostatistics, Iran; [3]Iran Shaheed Beheshti Medical University, Department of Biostatistics, Iran; [4]Shaheed Beheshti Medical University, Department of Health and Community Medicine

Correspondence to: Mohamad Amin Pourhoseingholi, Research Center for Gastroenterology and Liver disease, 7th floor of Taleghani Hospital, Tabnak Street, Evin, Tehran, Iran. E-mail: Amin_phg@yahoo.com, Aminphg@gmail.com

Abstract

Background: Logistic regression is one of the most widely used models to analyze the relation between one or more explanatory variables and a categorical response in the field of epidemiology, health and medicine. When there is strong correlation among explanatory variables, i.e. multicollinearity, the efficiency of model reduces considerably. The objective of this research was to employ latent variables to reduce the effect of multicollinearity in analysis of a case-control study about breast cancer risk factors.

Methods: The data belonged to a case-control study in which 300 women with breast cancer were compared to same number of controls. To assess the effect of multicollinearity, five highly correlated quantitative variables were selected. Ordinary logistic regression with collinear data was compared to two models contain latent variables were generated using either factor analysis or principal components analysis. Estimated standard errors of parameters were selected to compare the efficiency of models. We also conducted a simulation study in order to compare the efficiency of models with and without latent factors. All analyses were carried out using S-plus.

Results: Logistic regression based on five primary variables showed an unusual odds ratios for age at first pregnancy (OR=67960, 95%CI: 10184-453503) and for total length of breast feeding (OR=0). On the other hand the parameters estimated for logistic regression on latent variables generated by both factor analysis and principal components analysis were statistically significant (P<0.003). Their standard errors were smaller than that of ordinary logistic regression on original variables.

The simulation showed that in the case of normal error and 58% reliability the logistic regression based on latent variables is more efficient than that model for collinear variables.

Conclusions: This research indicated that logistic regression based on latent variables is more efficient than logistic regression based on original collinear variables.

Key words: multicollinearity, latent variables, factor analysis, principal components analysis, logistic regression, breast cancer

Introduction

Logistic regression is one of the most widely used generalized linear models to analyze the relation between one or more explanatory variables and a categorical response. 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 of explanatory variables of interest the efficiency of model reduces especially if there are also strong relationship among independent variables i.e. multicollinearity [1].

Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from, 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 in the first place.

So multicollinearity makes it difficult to apply directly the classical statistical methods to investigate relationships among variables and in some cases the prediction is felled in unexpected rang [2].

Let X is the matrix of explanatory variables that adjusted by centering on sample means and standard deviation or measurement scale. The correlation matrix will be $X^T X$.

The members of main diagonal of implies by

$$VIF_i = r_{ii} = \frac{1}{1-R^2_1} \qquad (1)$$

Called variance inflation factors (3) which is the multiple correlation of with other explanatory variables. Obviously as closes to one, the VIF increases.

Variance inflation factors (VIF) are a measure of the multicolinearity in a regression design matrix and a VIF more than 5 reflects a strong collinearity because VIF > 5 is the result of a linear correlation >0.8 [1].

What to do about multicollinearity? Increasing the sample size is a common first step since when sample size is increased, standard error decreases.

Another solution 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, which they should have been.

In some cases, variables involved in multicollinearity can be combined into a single variable that called latent variable. Latent variables are such kind of 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 [4].

In statistics, Latent variables (as opposed to observed variables), are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed and directly measured.

The using of latent variables for data reduction actually has had more application in psychology and social sciences [5], but because of some conditions in the areas of epidemiology and medical sciences which researchers often encounter situations where there are many variables related to each other, the problem of multicollinearity is expected to occur [6] and this technique can be done for this problem. However multicollinearity makes problem in logistic regression too, the researchers just have focused on linear regression with normal response.

The objective of this research was to employ latent variables to reduce the effect of multicollinearity in analysis of a case-control study about breast cancer risk factors and compare the efficiency of methods that generate latent variables with simulation.

Methods

1) Latent variable techniques

1.1) Principal components analysis

Principal components analysis is one of popular methods to data reduction in multivariate analysis. The history of this method is referred to Pearson's development in orthogonal least square fitting and the most theoretical development was related to Hotteling's study.

Principal components analysis is frequently used as a method for data reduction in linear regression to reduce the effect of multicollinearity [7].

In this technique the principal components are produced by eigen values matrix as a linear combination of original variables, Create new, uncorrelated variables, they are independent from each other and could replaced by original collinear variables in analysis [8].

Consider the following regression model:

$$y = X\beta + \varepsilon \ (i) \qquad (2)$$

Where y is a vector of η observation from a dependent variable and $X_{\eta x \rho}$ is a matrix of independent variables. β is a vector and regression coefficients and is a vector of random errors with variance equals^2.

Because the logistic regression, as a generalized linear model is an extension of linear regression based on logit transformation (9), principal component can be used in such model.

The value of principal components for each observation can be calculated by following

$$Z = XA \qquad (3)$$

Where the (i,k) 'th member of matrix Z is the k'th principal component for i'th observation from $A_{\rho x \rho}$that its k'th column is the k'th eigen vector of $X'X$.

Because of A's orthogonally, $X\beta$ can be expressed as $XAA'\beta = Z\gamma$ where $\gamma = A'B$.

Therefore the first regression equation is transformed to $y = Z\gamma + \varepsilon$ where the explanatory variables replaced by their principal components. In this study components were calculated using the covariance method.

1.2) Factor analysis

Factor analysis is another method to data reduction first introduced by Spearman with assuming a specific underlying model for the whole data; the independent factors are computed based on variance-covariance matrix or correlation coefficient matrix. [10] So the observed variables are modeled as linear combination of the factors, plus "error" terms.

This technique is frequently used to eliminate the multicollinearity in linear regression models [11]. However its history referred to psychological literatures.

Suppose γ is a response variable and x is a vector of ρ dimension explanatory variable.

γ can be a binary variable too.

If f is a vector of dim $k < \rho$ factor that computed based on vector x, we can estimate regression γ on f instead of γ on x. In this study factor analysis was employed, using maximum likelihood procedure in order to estimate latent factors.

Results

2.1) Data Example: Breast Cancer Risk Factors

Breast cancer is the most common type of cancer among women in the world (12), In Iran; breast cancer constitutes 21.4 % of all cancer cases reported in the country. The crude incidence rate of breast cancer in women is 22.4 per 100000 populations [13].

Yavari et al [14] carried out a hospital based case-control study to elucidate roles of reproductive factors in breast cancer. The total sample comprised of 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 live birth (AFLB), number of pregnancy (NP) and live birth (NLB), and total length of breast feeding (TLBF). All of the analysis was carried out using S-plus.

We consider these five reproductive variables because they show naturally a high linear correlation among themselves (Table 1).

We applied ordinary multiple logistic regression to these variables. From Table 2, we find that when we include all variables in the model only the estimate for AFP and AFLB are statistically significant, and all others are not significant with very large p-values. But the results for these two 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 make interpreting the effect that each of these variables has on the outcome very difficult. If we use the variance inflation factor as an index of multicollinearity, we have that NP is 4.57, NLB is 5.83, TLBF is 2.13, AFP is 15.41 and AFLB is 15.34 which are dramatically high for NLB, AFP and AFLB and moderate for NA and TLBF.

The correlation structure and what we know substantively about the observed variables suggest creating two latent variables.

Hence we consider the following latent variable model; factor analysis to generate two factors:

Factor1=
0.85NP+0.97NLB+0.68TLBF-0.26AFP-0.26AFLB
Factor2=
-0.21NP-0.23NLB-0.25TLBF+0.94AFP+0.95AFLB

And principal component analyses to generate two components:

Component1=
0.90NP+0.92NLB+0.82TLBF-0.25AFP-0.24AFLB
Component2=
-0.21NP-0.25NLP-0.22TLBF+0.96AFP+0.96AFLB

The value for the coefficient derived from both methods indicated that the first latent variable (Factor1 and Component1 in both techniques) has been weighted with NP, NLB and TLBF but the

Table 1. The Correlation matrix of the five reproductive variables and variance inflation factors.

(VIF)*	AFLB	AFP	TLBF	NLB	NP	
4.57	-0.42**	-0.44**	0.67**	**0.90	1	NP
5.83	-.047**	-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 1 percent level

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.

Then logistic regression has been conducted with these two latent variables. From Table 2, we see that for both latent variable methods the odds ratios of two latent variables are significant.

Then the response variable was generated as following:

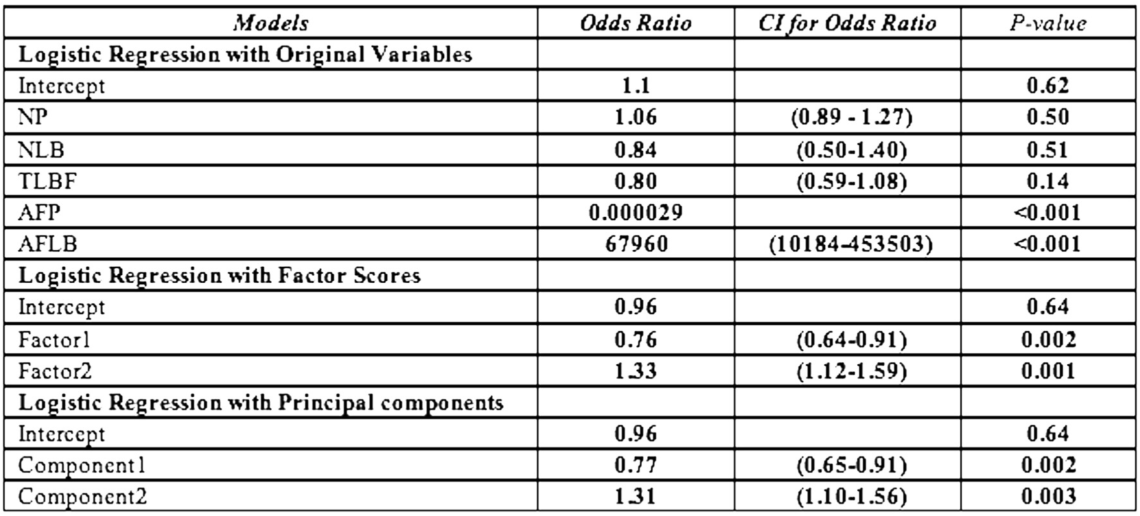
$$y = P\ (\ y\ -\ 1) = \frac{exp\ (\ 3 + 4g_1 + 2g_2)}{1 + exp\ (\ 3 + 4g_1 + 2g_2)} \qquad (5)$$

With respect to logit transformation:

$$y_2 = 1n\ (\ {}^{y_1}\!/_{1\ -\ y_1}) + ae_7 \qquad (6)$$

y_2 was generated as a logit response.

Table 2. Estimated coefficients of model using logistic regression with original variables and latent variables in Breast Cancer Risk Factors.



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**
Factor1	**0.76**	**(0.64-0.91)**	**0.002**
Factor2	**1.33**	**(1.12-1.59)**	**0.001**
Logistic Regression with Principal components			
Intercept	**0.96**		**0.64**
Component1	**0.77**	**(0.65-0.91)**	**0.002**
Component2	**1.31**	**(1.10-1.56)**	**0.003**

2.2) Simulation study:

We conducted a simulation study with an S-plus macro to examine whether these methods can perform well for eliminates multicollinearity in a logistic regression with high collinear explanatory variables and to evaluate that which latent variable model is the best.

We considered two uncorrelated variable groups consist of high correlated variables generated from the following model:

$$\begin{aligned} x_1 &= g_1 + 0.25e_1 \\ x_2 &= 5 + 0.5g_1 + 0.25e_2 \\ x_3 &= 3 + 0.7g_1 + 0.25e_3 \\ x_4 &= g_2 + 0.25e_4 \\ x_5 &= 3 + 0.7g_1 + 0.25e_5 \\ x_6 &= 2 + 0.8g_2 + 0.25e_6 \end{aligned} \qquad (4)$$

g_1&g_2 are independent from each other but have a same distribution. So the artificial variables calculated from them in two groups consist of x_1,x_2&x_3 generated from g_1 and x_4,x_5&x_6 generated from g_2 in second group are independent. The (ei,i = 1,...,7) reflects the random error term.

Two different distributional scenarios are considered for (g_1,g_2) and (ei,i = 1,...,7).

First these variables all are distributed i.i.d. $N(0,1)$, second to asses the effect of skewness the error terms distributed X^2 (1) with mean 0 and variance 1.We choose X^2 (1) because it represents a highly skewed distribution (Highly non-normal).

In each two distributional scenarios, two different levels for were selected (i.e. either 1 or 4) in order to control a high and low of reliability for the model as following:

$$Reliability = 1\ -\ {}^{Var\ (\ \varepsilon\ )}\!/_{Var\ (\ y\ )} \qquad (7)$$

The term of reliability is one minus the ratio of the variation of the error score and the variation of the observed score, which is how much the variation of the response variable represented by the model. The reliability decreases when the error term of the estimated model increases and vice versa [15]. Therefore in order to comparing the simulated results in different cases of reliability we set two different coefficients for error term. The reliability for equals 0.96 when and 0.58 when .

To evaluate the sample size variation on precision of estimation, each scenario was repeated with sample in 60, 300 and 600 size and finally we had 12 simulation scenarios (table 3).

For each of the training artificial data sets the simulation have been done of the 1000 iterations and two latent variables calculated by factor analysis and principal components analysis both and results were compared with empirical expected mean squared error from the following formula:

$$MSE_i = \Sigma_{j=1}^{N} \left(\hat{y} (\hat{o}_i , x_{ij}) - y_{ij} \right)^2 \Big/ n \qquad (8)$$

$$EMSE = \Sigma_{i=1}^{N} MSE_i \Big/ N \qquad (9)$$

Where $N= 1000$ is the number of iterations in simulation study and $n = 60,300$ and 600 for each scenario.

The final results of empirical expected mean squared error for all simulation scenarios are in table 3.

From figure 1 we see that efficiency of OLR is better than others in term of EMSE with high reliability and normal error, but with increasing in sample size, the efficiency of FS increases and goes up to OLR's and approximately they are both converged. The PC method performs slightly worse than both methods and there was no change in each sample size for its efficiency.

Figure 2 shows that the FS is the best with respect to low reliability and all the methods improved when sample size increases. The PC is better than OLR in n=60, 300.

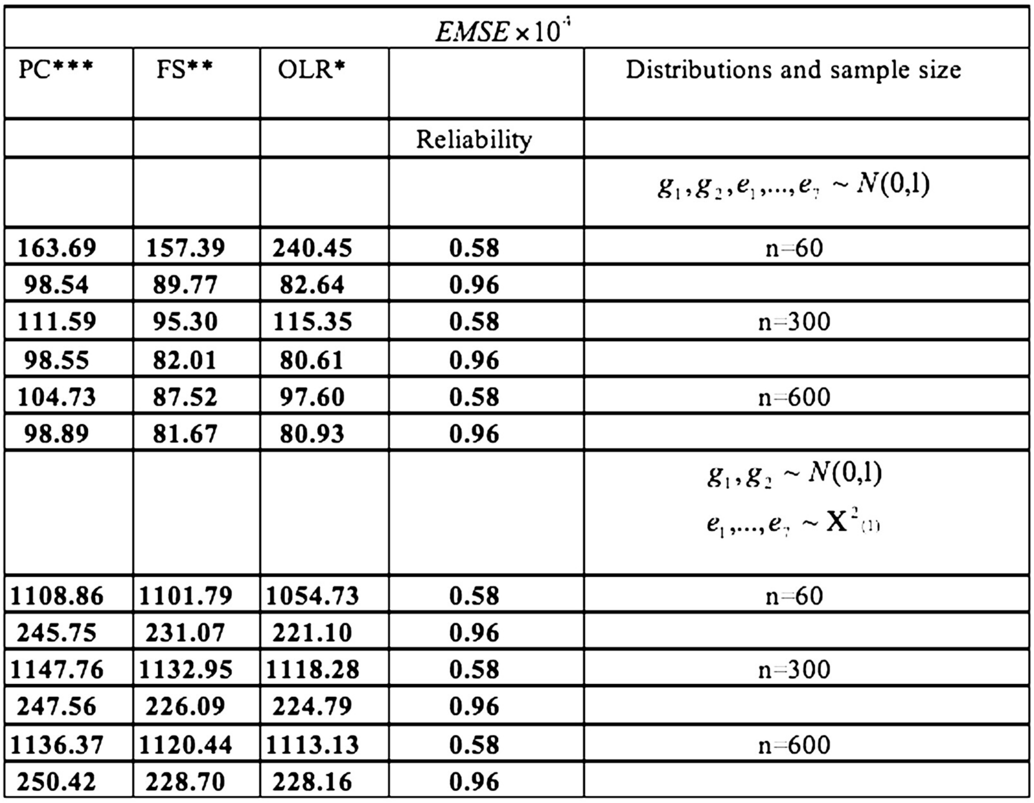
From figure 3 we find OLR the best but with increasing in sample size FS is converged to it.

Figure 4 showed that the OLR is the best too but there is no important difference among all methods in n=600

As is expected, the scale of the EMSE is larger when the reliability is lower because the error variance in the model is larger and each the methods are not robust to the non-normality when the error is Chi-square.

When n=600, there is practically no clear winner. Thus we see that the improvements of all methods approximately are closed to each other with increasing sample size.

Table 3. Empirical expected mean squared error for different sample size, reliability and errors.



$EMSE \times 10^{-4}$				
PC***	FS**	OLR*		Distributions and sample size
			Reliability	
				$g_1, g_2, e_1, \ldots, e_7 \sim N(0,1)$
163.69	**157.39**	**240.45**	**0.58**	n=60
98.54	**89.77**	**82.64**	**0.96**	
111.59	**95.30**	**115.35**	**0.58**	n=300
98.55	**82.01**	**80.61**	**0.96**	
104.73	**87.52**	**97.60**	**0.58**	n=600
98.89	**81.67**	**80.93**	**0.96**	
				$g_1, g_2 \sim N(0,1)$ $e_1, \ldots, e_7 \sim X^2_{(1)}$
1108.86	**1101.79**	**1054.73**	**0.58**	n=60
245.75	**231.07**	**221.10**	**0.96**	
1147.76	**1132.95**	**1118.28**	**0.58**	n=300
247.56	**226.09**	**224.79**	**0.96**	
1136.37	**1120.44**	**1113.13**	**0.58**	n=600
250.42	**228.70**	**228.16**	**0.96**	

*Ordinary Logistic Regression with Original Variables; **Logistic Regression with Factor Scores; *** Logistic Regression with Principal components

Figure 1. EMSE with Reliability= 0.96 and Normal error.

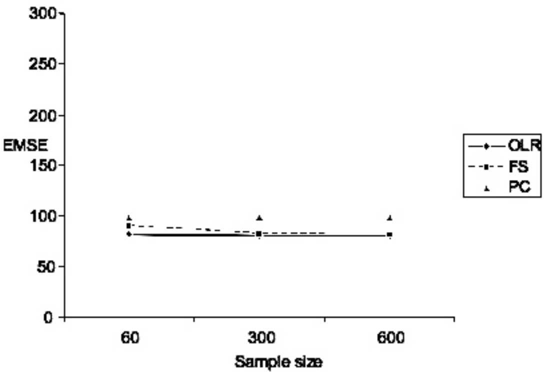


Figure 2. EMSE with Reliability= 0.58 and Normal error.

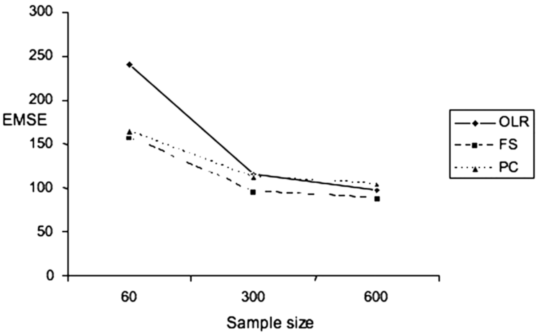
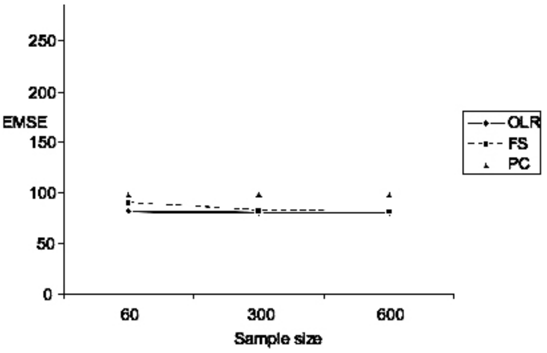


Figure 3. EMSE with Reliability= 0.96 and Chi-Square error.



Discussion

We have discussed using latent variables in a case-control study of breast cancer risk factors and we conducted a simulation which demonstrates that the use of latent variables in the case of normal error and low reliability yield a better fit and more efficiency for model and factor analysis is more efficient than principal components analysis.

Wall and Li [16] compared the linear regression of two methods of factor analysis based on generalized least square and structural equation modelling. Their results indicated a perfect fit for SEM in compare to FS. They also showed with real data, which was about respiratory disease, latent variable models produced parameters with smallest standard errors.

The skewness imposed a bad effect on efficacy. Especially this problem is touchable for FS because of using maximum likelihood to estimate the latent variables. However in large sample size the results were converged cross the all techniques. Wall and Li controlled the skewness in their own simulation. They used generalized least square, so the method was robust to the non normality. In order to evaluate the effect of sample size on their results they considered n=50, 200 and 1000 and found that the efficacy of FS increased with respect to increasing in sample size. These findings are acceptable with our results. But they didn't find similar results in the case of reliability equals 58% and 96% about FS method. On the other hand in our simulation FS is the best.

Morris [17] compared several different types of factor analysis with low reliability equals to 33% consistently found FS more favourable but its results seems sceptical because his simulation was based on re sampling of the same 1000 observations for all cases. Also Morris only examined the correlation between observations instead of using a statistic to compare the efficiency. Helland and Almoy [18] conducted a simulation study to compare principal component regression and partial least square regression. They found mixed results with no clear winner.

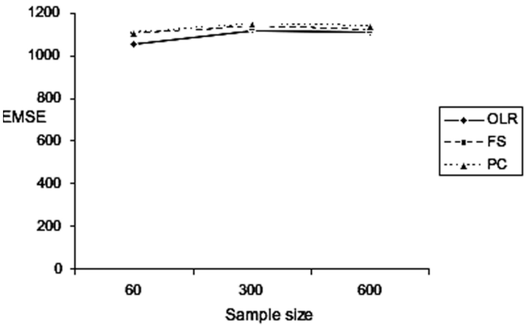
Filzmoser and Croux [19] suggested using principal component with a special algorithm in a linear regression with collinear explanatory variables. This algorithm has been used to choose the number of factors. They also by a simulation study proved that this method is efficient in a linear regression with high correlated variables.

Aguilera & Escabias [20] showed that PC is a favourable method in a logistic regression with collinear data. Escabias et al [21] used this method to modelling climatologically data.

The idea of using latent variables instead of original variables to data reduction comes from this fact that these variables can reflect the relation among observations [21], but the final comment should be kept in our mind about the benefits and restriction of these methods.

One of the biggest interests for any researcher is usually interpretation of parameters and what they exactly mean. A possible criticism of using the latent variable technique is the interpretation of the new parameter that calculated from new

Figure 4. EMSE with Reliability= 0.58 and Chi-Square error.



model. This interpretation is too difficult [22].

Somehow PC technique can be the first solution because principal components are a linear combination of initial data and opposite of FA do not assume an underlying model for dataset [8]. So researchers are able re estimate the parameters for original model, using inverse rotation matrix which translates the latent parameters to initial ones. Also some methods have been recommended to interpret the components too [23]. In spite of some disadvantages, the applications of latent variables are so increased that a new technique is proposed to produce latent variables for multi responses [24].

Acknowledgments

The authors wish to thanks Shaheed Beheshti Medical University for the support of breast cancer risk factor project.

References

1) Chattergee S, Hadi AS, Price B. Regression analysis by example. John Wiley & Sons, USA. 2002:225-58.
2) Myers RH. Classical and modern regression with applications. Pws-Kent Publishing Company. 1990:123-29.
3) Velleman P, Welsch R. Efficient computing of regression diagnostics. Am Statist 1981;35(4):234-42.
4) Van Eye A, Clogg CC. Latent variables analysis; application for developing research. SAGE publication. 1994:3-35.
5) Kenneth AB. Latent variables in psychology and the social sciences. Ann Rev Psychol 2002;53:605–34.
6) Kleinbaum D. Logistic Regression. New York: Springer, 1994:168.
7) Rawlings JO. Applied regression analysis:A research Tools. Wadsworth: Belmont, 1988:327-56.
8) Jollife IT. Principal component analysis. Springer, 1986:129-41.
9) Dobson A. An introduction to generalized linear models. USA: Chapman-Hall, 2001:115-31.
10) Srivastava MS. Methods of multivariate statistics. New York: John Wiley & Sons, 2002:397-50.
11) Lawley DN, Maxwell AE. Regression and factor analysis. Biometrika 1973;60:331-38.
12) Humple N, Jones S. "I don't really know, so it's a guess": women's breast cancer risk estimation." Asian Pacific J Cancer Prev 2004;5(4):428-32.
13) Shamsa AZ, Mohagheghi MA. Final report of the National project for cancer registry. Tehran: Islamic Republic of Iran. Cancer Institute, 2002.
14) Yavari P, Mousavizadeh M, Sadrol-Hafezi B, Mehrabi Y. Reproductive characteristics and the risk of breast cancer, A case-control study. Asian Pacific J Cancer Prev 2005;6:370-75.
15) Rudner LM, Shafer WD. Reliability. College Park, MD: ERIC Digest. ERIC Clearinghouse on Assessment and Evaluation, 2001.
16) Wall MM, Li R. Comparison of multiple regression to two latent variable techniques for estimation and prediction. Stat Med 2003;22:3671-85.
17) Morris J. The predictive accuracy of full-rank variables vs. various types of factor scores: implication for test validation. Educ Psychol Measurement 1980;40:389-96.
18) Helland IS, Almoy T. Comparison of prediction methods when only a few components are relevant. J Am Stat Ass 1994;89:583-91.
19) Filzmoser P, Croux C. A projection algorithm for regression with collinearity. In: Jajuga K, Sokolowski A, Bock HH, eds. Classification, clustering and data analysis. Springer, 2002.
20) Aguilera AM, Escabias M. Principal component logistic regression. Proceedings in computational statistics. Physica-Verlag, 2000:175-80.
21) Sobel ME. Causal inference in latent variable models. In: Van Eye A, Clogg CC eds. Latent variables analysis; application for developing research. SAGE publication, 1994:3-35.
22) Rencher AC. Methods of multivariate analysis. John Wiley & Sons, 2002.
23) Chipman HA, Gu H. Interpretable dimension reduction, 2002. Available from http: ace.acadiau.ca/math/chipmanh/publications.html [Accessed on april 2008].
24) Guo J, Wall MM, Amemiya Y. Latent class regression on latent factors. Biostat 2006;7:145 – 63.