COMPARING THREE METHODS OF HANDLING MULTICOLLINEARITY USING SIMULATION APPROACH

NORLIZA BINTI ADNAN

UNIVERSITI TEKNOLOGI MALAYSIA

COMPARING THREE METHODS OF HANDLING MULTICOLLINEARITY USING SIMULATION APPROACH

NORLIZA BINTI ADNAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

MEI 2006

To

My loving and supportive parents,

Hj Adnan and Pn Ramnah

My siblings,

Hairil, Aizam, Haffiz and Sarah

and

Pn Rusnah and family, and all my supportive friends

ACKNOWLEDGEMENT

"In the name od ALLAH the All-Merciful, The All-Compassionate. All praise be to ALLAH for giving me the strength and courage to complete this thesis"

A very great gratitude and appreciation expressed to all those who make a part to the successful cease of this thesis either directly or indirectly. Particularly, I wish to express my sincere appreciation to my main thesis supervisor, Dr. Maizah Hura binti Ahmad, for encouragement, guidance, critics and motivation. I am also very thankful to my co-supervisor Dr. Robiah binti Adnan for her guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here. An appreciation also goes to the Universiti Teknologi Malaysia for financial support.

Lots of gratitude and special thanks also to my family for their support, love and encouragement throughout my study. My sincere appreciation also extends to all my lovely friends.

ABSTRACT

In regression, the objective is to explain the variation in one or more response variables, by associating this variation with proportional variation in one or more explanatory variables. A frequent obstacle is that several of the explanatory variables will vary in rather similar ways. This phenomenon called multicollinearity, is a common problem in regression analysis. Handling multicollinearity problem in regression analysis is important because least squares estimations assume that predictor variables are not correlated with each other. The performances of ridge regression (RR), principal component regression (PCR) and partial least squares regression (PLSR) in handling multicollinearity problem in simulated data sets are compared to help and give future researchers a comprehensive view about the best procedure to handle multicollinearity problems. PCR is a combination of principal component analysis (PCA) and ordinary least squares regression (OLS) while PLSR is an approach similar to PCR because a component that can be used to reduce the number of variables need to be constructed. RR on the other hand is the modified least square method that allows a biased but more precise estimator. The algorithm is described and for the purpose of comparing the three methods, simulated data sets where the number of cases was less than the number of observations were used. The goal was to develop a linear equation that relates all the predictor variables to a response variable. For comparison purposes, mean square errors (MSE) were calculated. A Monte Carlo simulation study was used to evaluate the effectiveness of these three procedure. The analysis including all simulations and calculations were done using statistical package S-Plus 2000 software.

ABSTRAK

Objektif bagi regresi ialah untuk menerangkan variasi bagi satu atau lebih pembolehubah bersandar dengan cara menghubungkan variasi ini berkadaran dengan satu atau lebih pembolehubah tak bersandar. Halangan yang sering berlaku ialah apabila wujudnya kebersandaran antara pembolehubah-pembolehubah tak bersandar. Fenomena ini dipanggil multikolinearan. Mengawal dan mengatasi masalah multikolinearan di dalam analisis regresi adalah penting kerana kaedah penganggaran kuasa dua terkecil menganggap bahawa pembolehubah tak bersandar tidak berkorelasi antara satu sama lain. Perbandingan antara penggunaan kaedah regresi permatang (RR), regresi komponen berkepentingan (PCR) dan regresi sebahagian kuasa dua terkecil (PLSR) di dalam mengawal masalah multikolinearan dilakukan menggunakan set data yang disimulasi bagi membantu dan memberi satu pendekatan kepada para pengkaji yang akan datang tentang pemilihan kaedah terbaik bagi mengawal masalah multikolinearan. Kaedah regresi permatang ialah pengubahsuaian kaedah kuasa dua terkecil (LS) yang memasukkan pemalar kepincangan, δ di dalam penganggar kuasa dua terkecil. Regresi komponen berkepentingan pula merupakan gabungan analisis komponen berkepentingan (PCA) dengan kaedah kuasa dua terkecil biasa (OLS) sementara kaedah regresi sebahagian kuasa dua terkecil adalah hampir sama dengan kaedah regresi komponen berkepentingan di mana komponen baru perlu dibina untuk mengurangkan bilangan pembolehubah. Algoritma bagi setiap kaedah turut diterangkan dan untuk tujuan perbandingan bagi setiap kaedah, set data bagi kes bilangan pembolehubah tak bersandar lebih kecil dari bilangan pemerhatian. Perbandingan keberkesanan bagi ketiga-tiga kaedah tersebut menggunakan ralat min kuasa dua (MSE). Kaedah simulasi Monte Carlo digunakan untuk menilai keberkesanan ketiga-tiga kaedah yang dibincangkan. Semua simulasi dan pengiraan dilakukan dengan menggunakan pakej statistik S-PLUS 2000.

TABLE OF CONTENTS

2 LITERATURE REVIEW

3 METHODS FOR HANDLING MULTICOLLINEARITY

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

y	Response (dependent) variable
x	Predictor (independent) variable
β	Parameter (regression coefficient), known constant
ε	Error
σ^2	Variance
$\boldsymbol{Y}$	Matrix of observations
$\boldsymbol{X}$	Matrix of predictors
$\boldsymbol{\beta}$	Vector of parameters
$\boldsymbol{\varepsilon}$	Vector of error matrix term
$\boldsymbol{I}$	Identity matrix
$\hat{\beta}$	Estimate regression coefficient
$\hat{\boldsymbol{\beta}}$	Vector of estimate regression coefficient
$\hat{y}$	Fitted value
$\boldsymbol{H}$	Hat matrix
h_{ii}	*i*th diagonal element of hat matrix
e	Residual term
$\mathbf{e}$	Vector of residual term
n	Number of observations
p	Number of regressors
δ	Shrinkage parameter
$\boldsymbol{V}$	Matrix of normalized eigenvectors $\boldsymbol{X'X}$
$\boldsymbol{\Lambda}$	The diagonal matrix of eigenvalues of $\boldsymbol{X'X}$
T	Components for Partial Least Squares Regression

P	Matrix of x-loadings
$\mathbf{r}$	PLS weight vectors
k	Number of components for PLS and PCR
R^2	Coefficient of Determination
Z	Components for Principal Component Regression
λ	Eigenvalue

ABBREVIATIONS	MEANING
CI	Condition Indices
cov	Covariance
GCV	Generalized Cross Validation
LS	Least Squares
max	Maximum
MSE	Mean Square Error
OLS	Ordinary Least Squares
PCR	Principal Component Regression
PLSR	Partial Least Squares Regression
RMSE	Root Mean Square Error
RR	Ridge Regression
SMC	Squared Multiple Correlation
VIF	Variance Inflation Factors

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

Regression analysis is one of the most widely use of all statistical tools that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the others. The relationship of each predictor to the criterion is measured by the slope of the regression line of the criterion *Y* on the predictor. The regression coefficients are the values of these slopes.

The multiple linear regression model relates Y to $X_1, X_2, ..., X_p$ and can be expressed in terms of matrices as $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{y}$ is the *n*x1 vector of observed response values, $\boldsymbol{X}$ is the *n*x*p* matrix of *p* regressors, $\boldsymbol{\beta}$ is the *p*x1 regression coefficients and $\boldsymbol{\varepsilon}$ is the *n*x1 vector of error terms. The objectives of regression analysis are, (1) to find the estimates of unknown parameters $\boldsymbol{\beta}$'s and test of $\beta_1, \beta_2, ..., \beta_k$ for the significance of the associated predictors, (2) to use the regression equation to estimate *Y* from $X_1, X_2, ..., X_p$ and, (3) to measure the error involved in the estimation. The multiple

linear regression model may be used to identify important regressor variables that can be used to predict future values of the response variable.

The method of least squares is used to find the best line that on the average, is closest to all points. In other words, to find the the best estimates of β's with the least squares criterion which minimizes the sum of squared distances of all points from the actual observation to the regression surface. The name least squares comes from minimizing the squared residuals. From the Gauss-Markov theorem, least squares is always the best linear unbiased estimator (BLUE) and if ε is assumed to be normally distributed with mean 0 and variance σ^2, then least squares is the uniformly minimum variance unbiased estimator.

1.2 The Problem of Multicollinearity

In the applications of regression analysis, multicollinearity is a problem that always occurs when two or more predictor variables are correlated with each other. This problem can cause the value of the least squares estimated regression coefficients to be conditional upon the correlated predictor variables in the model. As defined by Bowerman and O'Connell (1990), multicollinearity is a problem in regression analysis when the independent variables in a regression model are intercorrelated or are dependent on each other.

There are a variety of informal and formal methods that have been developed for detecting the presence of serious multicollinearity. One of the most commonly used is the variance inflation factor (VIF) that measures how much the variances of the estimated regression coefficients are inflated compared to when the independent variables are not linearly related (Neter, et. al., 1990). The problem of multicollinearity can be remedied

using some method of estimation or some modifications of the method of least squares for estimating the regression coefficients.

The problem of multicollinearity can occur in both simple linear regression and multiple linear regression. Figure 1.1 illustrates the problem of multicollinearity that occur in simple regression (Wannacott and Wannacott, 1981). The figure shows how the estimate $\hat{\beta}$ becomes unreliable if the X_i's were closely bunched, that is, if the regressor X had little variation. When the X_i's are concentrated on one single value $\overline{X}$, then $\hat{\beta}$ is not determined at all. For each line, the sum of squared deviations is the same, since the deviations are measured vertically from $(\overline{X},\overline{Y})$. If $X_i = \overline{X}$, then all $x_i = 0$, and the term involving $\hat{\boldsymbol{\beta}}$ is zero. Hence, the sum of squares does not depend on $\hat{\boldsymbol{\beta}}$ at all. Therefore, when the values of $\boldsymbol{X}$ show little or no variation, then the effect of X on Y can no longer be sensibly investigated. The best fit for Y for data with multicollinearity was not a line, but rather a point $(\overline{X},\overline{Y})$. In explaining Y, multicollinearity makes the X_i's lose one dimension.

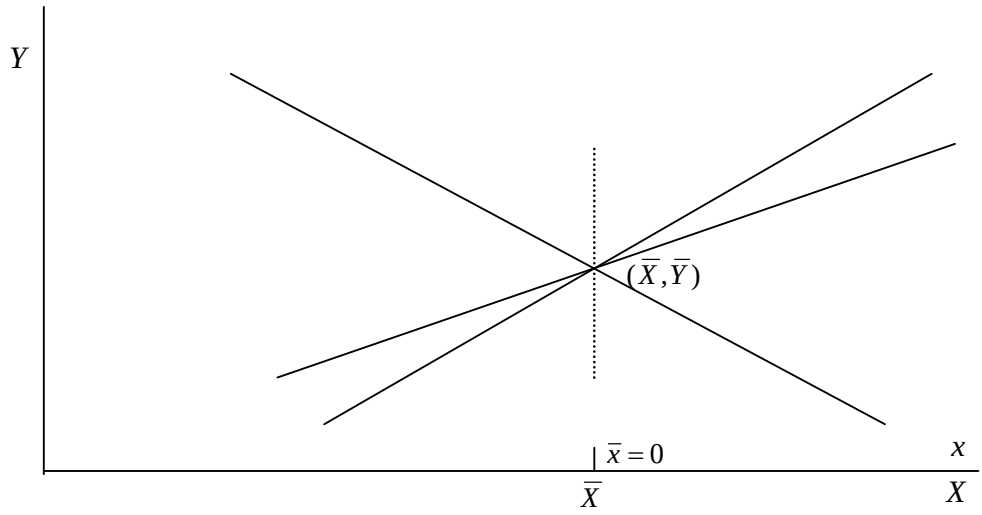

Figure 1.1 : Multicollinearity in simple regression

Figure 1.2 illustrates the problem of multicollinearity in multiple regression (Wannacott and Wannacott, 1981). All the observed points in the scatter plot lie in the vertical plane running up through *L*. In explaining *Y*, multicollinearity makes the X_i's lose one dimension and the best fit for *Y* is not a plane but instead the line *F*.

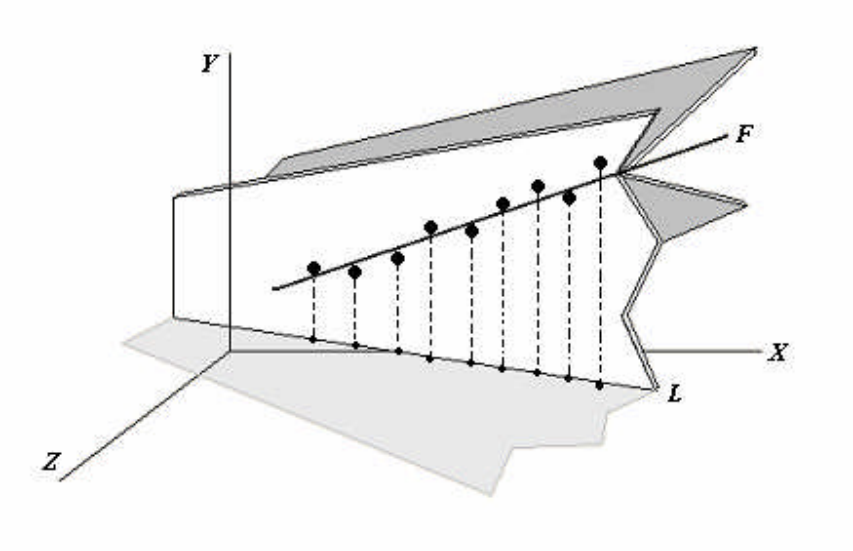

Figure 1.2 : Multicollinearity in multiple regression

Several approaches for handling multicollinearity problem have been developed such as Principal Component Regression, Partial Least Squares Regression and Ridge Regression. Principal Components Regression (PCR) is a combination of principal component regression analysis (PCA) and ordinary least squares regression (OLS). Partial Least Squares Regression (PLSR) is an approach similar to PCR because one needs to construct a component that can be used to reduce the number of variables. Ridge Regression is the modified least squares method that allow biased estimators of the regression coefficient.

1.3 Statement of Problem

This study will explore the following question :
Which method among Principal Component Regression, Partial Least Squares Regression and Ridge Regression performs best as a method for handling multicollinearity problem in regression analysis ?

1.4 Research Objectives

1. To compare the use of Partial Least Squares Regression, Principal Component Regression and Ridge Regression for handling multicollinearity problem.
2. To study the degree of efficiency among the three methods and hence rank them in terms of their capabilities in overcoming the multicollinearity problem using simulated data sets.

1.5 Scope of Research

For this research, the problem is focused on the analysis of multicollinearity problems in regression analysis using simulated data sets.

In Principal Component Regression, features from Principal Component Analysis (PCA) and multiple regression are combined to handle multicollinearity. The principal components computed a linear combinations of the explanatory variables (regressors). The Partial Least Squares Regression method is similar to Principal Component Regression because it is also a two-step procedure. However, PLSR searches for a set of

components called latent vectors that performs a simultaneous decomposition of *X* and *Y* with the constraint that those component explain as much as possible of the covariance between *X* and *Y*. This step generalizes PCA and it is followed by a regression step where the decomposition of *X* is used to predict *Y*.

The other method considered is the Ridge Regression. It is performed by adding a small biased estimator to the elements on the diagonal of the matrix $(\boldsymbol{X}'\boldsymbol{X})$ to be inverted, that is modifications of the least squares estimator.

1.6 Summary and Outline of Research

The goal of this research is to find the best procedure and method to handle multicollinearity problems by comparing the performances of the three methods to determine which method is superior than the others in terms of its practicality and efficiency. Practicality means how effective or convenient a method is in actual use while efficiency means how well the methods work in producing the best regression model and measured by a specified test discussed in Chapter 4. Basically three different methods are put forward, two of which are the methods with two-step procedures where it computes a component(s) and then regressed it on the response variable. The third method is a modification of the least squares method that allows biased estimators of the regression coefficients. The algorithms for each method used in this study are shown in Chapter 3.

Chapter 2 reviews the relevent literature on published work done recently concerning the problems of multicollinearity. Discussion on methods for handling multicollinearity problems in regression analysis is presented in Chapter 3. Chapter 4 describes the simulation work and the analysis of the three methods. Chapter 5 discusses the performances of the three methods and makes comparisons among them. Lastly, Chapter 6 concludes the study and makes recommendations for further research.

variable. The advantages of PCR are that hypothesis testing can be performed, and that complete optimisation is used in determining the PCs.

6.4 Future Research

The performance of the three methods can also be done by comparing the use of all methods for high-dimensional regressors where $p > n$. It is known that the problem of multicollinearity is present in the data set where the number of variables is high compared to the number of observations.

These three methods can also be considered in the handling of multiple outliers problem since according to Engelen et al. (2003), PLS and PCR are very sensitive to the presence of outliers in data set. The robust version of PLS and PCR are known to resist several types of contamination.

REFERENCES

Abdi, H., (2003a). Least Squares. In M. Lewis-Beck, A. Bryman, T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks (CA): Sage.

Abdi, H., (2003b). Partial Least Squares Regression (PLS regression). In M. Lewis-Beck, A. Bryman, T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks (CA): Sage.

Afifi, A.A. and Clark, Y., (1984). *Computer-aided multivariate analysis.* Lifetime Learning Publications, Belmont, California.

Barker, M., (1997). *A Comparisons of Principal Component Regression and Partial Least Squares Regression*. Multivariate Project.

Björkström, A., (2001). *Ridge Regression and Inverse Problem*. Research Report. Stockholm University, Sweden.

Björkström, A. and Sundberg, R., (1999). A Generalized View on Continuum Regression. *Scandinavian Jornal Statistics.* **25** : 17 – 30.

Bowerman, B.L. and O'Connell, R.T. (1990). *Linear Statistical Models an Applied Approach*. 2nd Ed., Boston : PWS-KENT Publishing Co.

Boeris, M.S., Luco, J.M. and Olsina, R.A., (2000). Simultaneous Spectrophotometric Determination of Phenobarbital, Phenytoin and Methylphenobarbital in Pharmaceutical Preparations by Using Partial Least Squares and Principal Component Regression Multivariate Calibration. *Journal of Phamaceutical and Biomedical Analysis.* **24** : 259 – 271.

De Jong, S. (1993), SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems.* **18** : 251 – 263.

De Jong, S., Farebrother and R.W., (1994). Extending the Relationship Between Ridge Regression and Continuum Regression. *Chemometrics and Intelligent Laboratory Systems*. **25** : 179 – 181.

Dempster, A.P., Schatzoff, M. and Wermuth, N., (1977). A Simulation Study of Alternatives to Ordinary least Squares. *Journal of the American Statistical Association*. **72** : 77 – 91.

Dijkstra, T. (1983), Some Comments on Maximum Likelihood and Partial Least Squares Methods. *Journal of Econometrics*. **22** : 67 – 90.

Dijkstra, T. (1985), *Latent Variables in Linear Stochastic Models: Reflections on Maximum Likelihood and Partial Least Squares Methods.* Second Edition, Amsterdam, The Netherlands: Sociometric Research Foundation.

Engelen, S., Hubert, M., Vanden B.K. and Verboven, S. (2003). Robust PCR and Robust PLSR: a comparative study. *Theory and Applications of Recent Robust Methods*. edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Series : Statistics for Industry and Technology, Birkhauser, Basel.

Farebrother, R. W., (1999). A Class of Statistical Estimators Related to Principal Components. *Linear Algebra and its Aplications*. **289 :** 121 – 126.

Filzmoser, P., Croux, C., (2002). A Projection Algorithm for Regression with Collinearity. *Classification, Clustering and Data Analysis*, 227 – 234.

Filzmoser, P., Croux, C., (2003). Dimension Reduction of the Explanatory Variables in Multiple Linear Regression. *Pliska Studia Mathematica Bulgaria.* **14** : 59 – 70.

Garthwaite, P.H., (1994). An Interpretation of Partial Least Squares. *Journal of the American Statistical Association*. **89** : 122 – 127.

Geladi, P., Kowalski, B.R., (1986a). Partial Least-Squares Regression: A Tutorial. *Analytica Chimica Acta*. **185** : 1 – 17.

Geladi, P., Kowalski, B.R., (1986b). An Example of 2-Block Predictive Partial Least Squares Regression with Simulated Data. *Analytica Chimica Acta*. **185** : 19 – 32.

Gibbons, D.G., (1981). A Simulation Study of Some Ridge Estimators. *Journal of the American Statistical Association.* **76** : 131 – 139.

Gruber, M.H.J., (1990). *Regression Estimators : A Comparative Study.* California : Academic Press

Hair, J. F., Anderson, R.E., Tatham, R.L. and Black, W.C., (1998). *Multivariate Data Analysis.* 5th Ed., New Jersey : Prentice Hall.

Hansen, P.M. and Schjoerring, J.K., (2003). Reflectance Measurement of Canopy Biomass and Nitrogen Status in Wheat Crops Using Normalized Difference Vegetation Indices and Partial Least Squares. *Remote Sensing of Environment.* **86** : 542 – 553.

Hawkins, D.M. and Yin, X., (2002). A Faster Algorithm for Ridge Regression of Reduced Rank Data. *Journal of Computional Statistics & Data Analysis*. **40** : 253 – 262.

Hocking, R.R, (1996). *Methods and Applications of Linear Models : Regression and the Analysis of Variance.* USA : John Wiley & Sons.

Hoerl, R.W., Schuenemeyer, J.H. and Hoerl, A.E., (1986). A Simulation of Biased Estimation and Subset Selection Regression Techniques. *Technometrics*. **28(4)** : 369 – 380.

Hoerl, A.E., Kennard, R.W., (1970). Ridge Regression : Biased Estimation to Nonorthogonal Problems. *Technometrics.* **12** : 56 – 67.

Hubert, M., Branden .K.V., (2003a). Robust methods for Partial Least Squares Regression, *Journal of Chemometrics*. **17** : 537-549.

Hubert, M., Branden .K.V., (2003b). Robustness properties of a robust Partial Least Squares Regression, *Journal of Chemometrics***. 17** : 537-549.

Hwang, J.T., Nettleton, D., (2000). Principal Components Regression With Data-Chosen Components and Related Methods. *Technometrics*. **45** : 70 – 79.

Jennrich, R.I., (1995). *An Introduction to Computational Statistics : Regression Anaysis.* New Jersey : Prentice Hall.

Kvalheim, O. M. (1987). Latent-structure decompositions (projections) of multivariate data. *Chemometrics and Intelligent Laboratory Systems*. **2** : 283 – 290

Kleinbaum, D.G., Kupper, L.L., Muller, K.E. and Nizam, A., (1998). *Applied Regression Analysis and Other Multivariable Methods.* USA : DUXBURY Press.

Li, B., Martin, E.B. and Moris, A.J., (2001). Box-Tidwell Transformation Based Partial Least Squares Regression. *Computers and Chemical Engineering.* **25** : 1219 – 1233.

Marquardt, (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation. *Technometrics*. **12** : 591 – 612.

Miller, A.J., (1990). *Subset Selection in Regression*. Chapman & Hall, New York.

Myers, R.H., **(1986).** *Classical and Modern Regression With Applications.* 2nd Ed. USA : PWS-KENT Publishing Company.

Neter, J., Wasserman, W. and Kutner, M.H., (1985). *Applied Linear Statistical Models*. 2nd Ed. USA : IRWIN.

Neter, J., Wasserman, W. and Kutner, M.H., (1990). *Applied Linear Regression Models*. 3rd Ed. USA : IRWIN Book Team.

Neter, J., Kutner, M.H., Nachtseim, C.J. and Wasserman, W., (1996). *Applied Linear Statistical Models.* 4th Ed. USA : Irwin Bokk Team.

Ozcelik, Y., Kulaksiz S. and Cetin, M.C., (2002). Assessment of the Wear of Diamond Beads in the Cutting of Different Rock Types by the Ridge Regression. *Journal of Materials Processing Technology*. **127** : 392 – 400.

Rougoor, C.W., Sundaram, R. and Van Arendonk, J.A.M., (2000). The Relation Between Breeding Management and 305-day Milk Production, Determined via Principal Components Regression and Partial Least Squares. *Livestock Product Science*.
66 : 71 – 83.

Serneels, S. and Van Espen, P.J., (2003). Sample specific prediction intervals in SIMPLS. In *PLS and related methods*, Ed. Vilares, M., Tenenhaus, M., Coelho, P., Esposito, V., Vinzi and Morineau, A., DECISIA, Levallois Perret (France), 219 – 233.

Stone, M. and Brooks, R.J., (1990). Continuum Regression : Cross-Validated Sequentially Constructed Prediction Embracing OLS, PLS and PCR. *Journal of Royal Statistic Society.* **52** : 237 – 269.

Sundberg, R., **(1993).** Continuum Regression and Ridge Regression. *Journal of Royal Statistics Society*. **55** : 653 – 659.

Sundberg, R., **(2002).** Continuum Regression. Article for 2nd Ed. of *Encyclopedia of Statistical Sciences*.

Tobias, R.D., (1997). *An Introduction to Partial Least Squares Regression*. Cary, NC : SAS Institute.

Wannacott, T.H. and Wannacott, R.J., (1981). *Regression :A Second Course in Statistics.* USA : John Wiley & Sons.

Wesolowsky, G.O., (1976). *Multiple Regression and Analysis of Variance.* USA : John Wiley & Sons.

Wold, H. (1966), Estimation of Principal Components and Related Models by Iterative Least Squares. In *Multivariate Analysis*, Ed. Krishnaiah, P.R., New York : Academic Press, 391 – 420.

Wold, S. (1994), PLS for Multivariate Linear Modeling. *QSAR: Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry*, ed. H. van de Waterbeemd, Weinheim, Germany: Verlag-Chemie.

Xie, Y.L. and Kalivas, J.H., (1997a). Evaluation of Principal Component Selection Methods to form a Global Prediction Model by Principal Component Regression. *Analytica Chimica Acta.* **348** : 19 – 27.

Xie, Y.L. and Kalivas, J.H., (1997b). Local Prediction Models by Principal Component Regression. *Analytica Chimica Acta.* **348** : 29 – 38.

Younger, M.S., (1979). *A Handbook for Linear Regression.* USA : DUXBURY Press.