Exploring the Performance of Methods to Deal Multicollinearity: Simulation and Real Data in Radiation Epidemiology Area

Mickaël Dubocq[1,2,3], Nadia Haddy[1,2,3], Boris Schwartz[1,2,3], Carole Rubino[1,2,3], Florent Dayet[1,2,3], Florent de Vathaire[1,2,3], Ibrahima Diallo[1,2,3] and Rodrigue S. Allodji[1,2,3,*]

[1]*Radiation Epidemiology Group, INSERM U1018, Villejuif, F-94805, France*

[2]*Gustave Roussy, Villejuif, F-94805, France*

[3]*Univ. Paris-Sud, Villejuif, F-94800, France*

Abstract: The issue of multicollinearity has long been acknowledged in statistical modelling; however, it is often untreated in the most of published papers. Indeed, the use of methods for multicollinearity correction is still scarce. One important reason is that despite many proposed methods, little is known about their strength or performance. We compare the statistical properties and performance of four main techniques to correct multicollinearity, i.e., Ridge Regression (R-R), Principal Components Regression (PC-R), Partial Least Squares Regression (PLS-R), and Lasso Regression (L-R), in both a simulation study and two real data examples used for modelling volumes of heart and Thyroid as a function of clinical and anthropometric parameters. We find that when the statistical approaches were used to address different levels of collinearity, we observed that R-R, PC-R and PLS-R appeared to have a somewhat similar behavior, with a slight advantage for the PLS-R. Indeed, in all implemented cases, the PLS-R always provided the smallest value of root mean square error (RMSE). When the degree of collinearity was moderate, low or very low, the L-R method had also somewhat similar performance to other methods. Furthermore, correction methods allowed us to provide stable and trustworthy parameter estimates for predictors in the modelling of heart and Thyroid volumes. Therefore, this work will contribute to highlighting performances of methods used only for situations ranging from low to very high multicollinearity.

Keywords: Lasso Regression, Multicollinearity, Organs volume modelling, Partial Least Squares Regression, Principal Components Regression, Ridge Regression.

1. INTRODUCTION

The issue of multicollinearity is very common in many research areas [1]. It has long been acknowledged in statistical modelling; however, it is often untreated in analyses and in the most of published papers [2]. Typically, multicollinearity is a problem in multiple regressions that develops when one or more of the explanatory variables are highly correlated with one or more of the other explanatory variables. If it is regressed on the other explanatory variables, then the matrix of intercorrelations among the explanatory variables is singular and there exists no unique solution for the regression coefficients [3]. Consequently, regression coefficients biased by multicollinearity might cause variables that demonstrate no significant relationship with the outcome when considered in isolation to become highly significant in conjunction with collinear variables, yielding an elevated risk of false-positive results (Type I error). Alternatively, multiple regression coefficients might show no statistical significance due to incorrectly estimated wide confidence intervals, yielding an elevated risk of false-negative results (Type II error). A broad variety of methods for multicollinearity correction have been developed [4], but these methods have rarely been applied possibly because their ability to correct multicollinearity is poorly understood.

The aim of this paper was to compare the statistical properties and performance of four main techniques to correct multicollinearity, i.e., the Ridge Regression method denoted R-R, the Principal Components Regression method denoted PC-R, the Partial Least Squares Regression method denoted PLS-R, and the Lasso Regression method denoted L-R. Comparisons were performed using a simulation study and within two datasets used for modelling organs volume as a function of clinical and anthropometric parameters.

2. MULTIPLE LINEAR REGRESSIONS AND MULTICOLLINEARITY EFFECTS

2.1. Multiple Linear Regression

Assume that there are **N** observations ($y, x_1, ..., x_k$), and the purpose is to build a predictor for the scalar dependent variable or response ***y*** based on the ***K***-dimensional vector **x** of regressors. Say that **x** is easier or cheaper to measure than **y**. The data used for regression can be collected in the matrix **X** and the vector **y**. Assume that the relationship between **X** and **y**

*Address correspondence to this author at the Radiation Epidemiology Group, INSERM U1018, Villejuif, F-94805, France; Tel: +33-014-211-5498; Fax: +33-014-211-5315; E-mail: rodrigue.allodji@gustaveroussy.fr

E-ISSN: 1929-6029/18

is linear. Without loss of generality, we assume that **X** is centred. The model can then be written as:

$$y = \mathbf{1}\beta_0 + \mathbf{X}\beta + \mathrm{e} \quad (1)$$

for the residual vector e.

The main problem is to estimate the regression vector **b** in order to obtain a predictor

$$\hat{y} = \bar{y} + \mathbf{X}^t\hat{\beta} \quad (2)$$

which gives as good predictions of unknown ***y'***s as possible. A measure of prediction accuracy, which is much used, is mean square error (MSE) defined by:

$$\mathrm{MSE}(\hat{y}) = \mathrm{E}(\hat{\mathrm{y}} - \mathrm{y})^2 \quad (3)$$

The most frequently used method of estimation for the regression vector is least squares [4]. The sum of squared residuals (RSS) is minimised over the space of **b** values. If we assume that **X** has rank ***k***, then so does $\mathbf{X}^t\mathbf{X}$ and so $\mathbf{X}^t\mathbf{X}$ is invertible and the least squares estimator has a very nice closed form solution:

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathrm{y} \quad (4)$$

The covariance matrix of β is equal to:

$$\mathbf{COV}(\hat{\beta}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} \quad (5)$$

This can also be written as follows:

$$\mathbf{COV}(\hat{\beta}) = \sigma^2\sum_{k=1}^{K}\mathbf{p}_k(1/\lambda_k)\mathbf{p}_k^t \quad (6)$$

where the ***p'***s are the eigenvectors of $\mathbf{X}^t\mathbf{X}$ and the λ's are the corresponding eigenvalues.

2.2. Multicollinearity Effects

A common situation in many applications of linear models is that there are highly correlated among the **x**-variables. When the linear relations are exact, the inverse of $\mathbf{X}^t\mathbf{X}$ does not exist and no unique $\hat{\mathbf{b}}$ can be produced. This problem is a form of ill conditioning in the ($\mathbf{X}^t\mathbf{X}$) matrix. If there is at least one near dependency in the data, one or more of the eigenvalues will be small. This implies that there are near dependencies among the columns of **X** [5]. It is well known that multicollinearity tends to produce least squares estimates that are too large in absolute value and whose signs may actually reverse [6]. While the method of least squares will generally produce poor estimates of the individual model parameters when strong multicollinearity is present, the estimated coefficients are correlated (confounding) with each other. Along with this correlation, multicollinearity has a multitude of other ramifications on our analysis, including: inaccurate regression coefficient estimates (leading to a high MSE), inflated standard errors of the regression coefficient estimates, deflated t-tests for significance testing of the regression coefficients, false non significance determined by the p-values, and degradation of model predictability.

3. METHODS FOR CORRECTING MULTICOLLINEARITY EFFECTS

3.1. Ridge Regression (R-R)

Hoerl and Kennard have suggested the Ridge regression as an alternative procedure to the method of least squares in regression analysis, especially when multicollinearity exists [7]. The Ridge regression was originally developed to invert the matrix $\mathbf{X}^t\mathbf{X}$ and is based on adding a positive scalar κ to the diagonal elements of κ. If κ is not null, the new matrix $\mathbf{X}^t\mathbf{X}+\kappa\mathbf{I}$ will be invertible. By using an improved least square method, ridge regression sought standardized coefficients. κ is called ridge parameter, and usually $0 < \kappa < 1$. κ value was selected when all the regression coefficients were relatively stable and the sign of the coefficients did not change. The ridge estimator, $\hat{\beta}_{\mathrm{Ridge}}$ is defined as follows:

$$\hat{\beta}_{\mathrm{Ridge}}(\kappa) = (\mathbf{X}^t\mathbf{X}+\kappa\mathbf{I})^{-1}\mathbf{X}^t\mathrm{y} \quad (7)$$

where κ is a positive scalar and **I** an identity matrix. If $\kappa = 0$, the ridge estimator become as the least squares [8]. An appropriate value of κ may be determinate by examination of the ridge trace, of the variance inflation factor (VIF) and of the RMSE. The ridge trace is a plot of the elements of $\hat{\beta}_{\mathrm{Ridge}}$ as function of κ [9]. Further details on the efficient choice of biasing constant for Ridge Regression are given in several papers [10]. The SAS regression procedure 'proc reg' can be told to generate a ridge trace, so that ridge regression is easy to implement [11].

3.2. Principal Components Regression (PC-R)

One of the simplest ways that the collinearity problem is solved in practice is by the use of principal component regression (PC-R). In PC-R instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used. This method usually gives much better results than least squares for prediction purposes when used successfully [12]. With this method, the original *k* explanatory variables are transformed into a new set of orthogonal or uncorrelated variables called principal components of

the correlation matrix. This transformation ranks the new orthogonal variables in order of their importance and the procedure then involves eliminating some of the principal components to effect a reduction in variance. After elimination of the least important principal components, a multiple regression analysis of the response variable against the reduced set of principal components is performed using ordinary least squares estimation. Because the principal components are orthogonal, they are pair-wise explanatory and hence ordinary least squares is appropriate. Once the regression coefficients for the reduced set of orthogonal variables have been calculated, they are mathematically transformed into a new set of coefficients that correspond to the original or initial correlated set of variables. These new coefficients are principal component estimators. The computational technique of PC-R method may be summarized as follows. Let X be the centered scaled $n \times k$ data matrix as given in equation (1). The $k \times k$ correlation matrix of the explanatory variables is then $\mathbf{C} = \mathbf{X}^t\mathbf{X}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the eigenvalues of the correlation matrix, and $\mathbf{V} = \begin{bmatrix} \mathrm{v}_1 & \mathrm{v}_2 & \ldots & \mathrm{v}_k \end{bmatrix}$ be the $k \times k$ matrix consisting of the normalized eigenvectors associated with each eigenvalue. The eigenvectors are the solutions of the determinant equation $\left|\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I}\right| = 0$, and associated with each eigenvalue, λ_j, is a vector, v_j, that satisfies the set of homogeneous equations $\left(\mathbf{X}^t\mathbf{X} + \lambda_j\mathbf{I}\right)\mathrm{v}_j = 0$. The vectors, $\mathbf{v}_j = \begin{pmatrix} v_{1j} & v_{2j} & \ldots & v_{kj} \end{pmatrix}^t$ are orthogonal to one another, hence $\mathbf{V}$ is orthonormal, where $\mathbf{V}^t\mathrm{V} = \mathrm{I}$ that is the identity matrix. Now consider the model formulation given in equation (1), that is, $y = \mathbf{1}\beta_0 + \mathbf{X}\beta + \mathrm{e}$. One can write the original regression model (equation 1) in the form:

$$y = \mathbf{1}\beta_0 + \mathbf{X}(\mathbf{V}^t\mathbf{V})\beta + \mathrm{e} \tag{8}$$

or

$$y = \mathbf{1}\beta_0 + Z\alpha + \mathrm{e} \tag{9}$$

where $\mathbf{Z} = \mathbf{X}\mathrm{V}$ and $\alpha = \mathrm{V}^t\beta$. $\mathbf{Z}$ is an $n \times k$ data matrix of principal components and $\alpha = \begin{pmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_k \end{pmatrix}^t$ is an $n \times 1$ vector of new coefficients. The model formulation in equation (9) is nothing more than the regression of the response variable on the principal components, and the transformed data matrix $\mathbf{Z}$ consists of the *k* principal components. Therefore, if the response variable ($\mathbf{y}$) is regressed against these *k* principal components using the model in equation (12), then the least squares estimator for the regression coefficients in vector α is the vector:

$$\hat{\alpha} = \left(Z^tZ\right)^{-1}Z^ty \tag{10}$$

and the variance-covariance matrix of the estimated coefficients in vector $\hat{\alpha}$ is given by:

$$Var(\hat{\alpha}) = \hat{\sigma}^2\left(Z^tZ\right)^{-1} = \hat{\sigma}^2 diag\left(\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_k^{-1}\right) \tag{11}$$

Even though the new variables are orthogonal, the same magnitude of variance is retained. But if multicollinearity is severe, there will be at least one small eigenvalue. An elimination of one or more principal components associated with the smallest eigenvalues will reduce the total variance in the model and thus produce an appreciably improved diagnostic or prediction model.

Suppose with *k* variables and hence *k* principal components, $r < k$ components are eliminated. From equation (9), with the retention of all components, $\alpha = \mathrm{V}^t\beta$, and the coefficients for the centered and scaled explanatory variables are obtained as:

$$\beta_{\mathrm{PC-R}} = \mathrm{V}\alpha \tag{12}$$

The PC-R can be implemented in SAS relatively easily using PRINCOMP procedure [11].

3.3. Partial Least Squares Regression (PLS-R)

Partial Least Squares (PLS) is a method for constructing predictive models when the variables are too many and highly collinear. Besides collinearity, PLS-R is also robust against other data structural problems such as skew distributions and omission of regressors [13]. Like principal component analysis, the basic idea of PLS is to extract several latent factors and responses from a large number of observed variables [14]. More specifically, the aim is to predict the response by a model that is $\mathbf{C\hat{O}V(Y,F)} = \frac{1}{\mathrm{n}}\mathbf{v}^{\mathrm{t}}\mathbf{X}^{\mathrm{t}}\mathbf{Y}$ based on linear transformations of the explanatory variables. Therefore, the acronym PLS is also taken to mean Projection to Latent structure [13]. The latent factors are then used for prediction in place of the original variables [15]. In order to specify the latent component matrix $\mathbf{F}$ such that $\mathbf{F} = \mathbf{XV}$, PLS-R requires finding the columns of $\mathbf{V} = \begin{bmatrix} \mathrm{v}_1 & \mathrm{v}_2 & \ldots & \mathrm{v}_k \end{bmatrix}$ from successive optimization problems. The sample covariance between the response variable $\mathbf{Y}$ and the random variable $\mathbf{F} = \mathbf{v}_1\mathbf{X}_1 + \cdots + \mathbf{v}_{\mathrm{k}}\mathbf{X}_{\mathrm{k}}$ can be computed as

$$\mathbf{C\hat{O}V(Y,F)} = \frac{1}{\mathrm{n}}\mathbf{v}^{\mathrm{t}}\mathbf{X}^{\mathrm{t}}\mathbf{Y} \tag{13}$$

since the matrices $\mathbf{X}$ and $\mathbf{Y}$ contain the centered data. Similarly, for the sample variance of the random variable $\mathbf{F}$, we have

$$\mathbf{\hat{V}ar(F)} = \mathbf{v}^t\mathbf{X}^t\mathbf{X}\mathbf{v} = \frac{1}{n}\mathbf{v}^t\mathrm{v} \quad (14)$$

The criterion to find the k^{th} direction vector v_k for univariate $\mathbf{Y}$ is formulated as

$$v_k = arg\,max_v\, cor^2(Y, Xv)\, var(Xv), \quad (15)$$

where $\mathbf{v}^t\mathrm{v} = 1$, for $j = 1,\ldots,k-1$.

As evident from this formulation, PLS-R seeks direction vectors that not only relate X to Y but also capture the most variable directions in the X space [13]. The maximal number of such latent factors that have nonzero covariance with $\mathbf{Y}$ is

$$C_{max} = min(n-1, k). \quad (16)$$

The weight vectors $\mathrm{v}_1\ \mathrm{v}_2\ \ldots\ \mathrm{v}_k$ can be computed sequentially via a simple and fast non-iterative algorithm given, e.g. in and denoted as 'algorithm with orthogonal scores' because the matrix $\mathbf{F}^t\mathbf{F}$ is diagonal [16]. Finally, the matrix $\mathbf{B}$ of regression coefficients for the model in equation (1) is given as

$$\mathrm{B} = \mathrm{V}(\mathrm{F}^t\mathrm{F})^{-1}\mathrm{F}^t\mathrm{Y}. \quad (17)$$

It can be shown that the resulting regression coefficients in matrix $\mathbf{B}$ are the same with both algorithms [16]. The PLS-R can be implemented in SAS relatively easily using PLS procedure [11].

3.4. Least Absolute Shrinkage and Selection Operator (Lasso) Regression

The Lasso is a form of regularized or "penalized" regression proposed by Tibshirani [17]. When there are high correlations between predictors, lasso is useful as R-R, PC-R and PLS-R. This method allows to obtain an estimator of regression coefficient which minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant $\delta \geq 0$. When δ is small, some regression coefficients will be necessarily to zero or very close to zero, else the sum of the regression coefficients will exceed δ value. In view of shrinking the regression coefficients by imposing a penalty on their size, the Lasso is similar in spirit to Ridge regression. However, ridge regression cannot produce a parsimonious model, as it always keeps all the predictors in the model [18].

Without loss of generality, throughout this article, we assumed that the data were standardized to have mean 0. That was, $\mathbf{1}^t\mathrm{y} = 0$, $\mathbf{1}^t\mathrm{x}_j = 0$, and $\mathrm{x}_j{}^t\mathrm{x}_j = 1$ for $j = 1,\ldots,k$. The Lasso estimate was the solution to

$$\mathbf{min}_{\beta}(\mathrm{y} - \mathrm{X}\beta)^t(\mathrm{y} - \mathrm{X}\beta), \text{ subject to } \sum_{j=1}^{k}|\beta_j| \leq \delta. \quad (18)$$

Tibshirani noted that the lasso constraint $\sum_{j=1}^{k}|\beta_j| \leq \delta$ was equivalent to adding the penalty term $\tau\sum_{j=1}^{k}|\beta_j|$ to the residual sum of squares, so there was a direct correspondence between parameters $\delta \geq 0$ and $\tau \in [0, +\infty)$ [17]. Thus, an alternative formulation of the Lasso was defined by

$$\mathbf{min}_{\beta}\frac{\mathbf{1}}{n}(\mathrm{y} - \mathrm{X}\beta)^t(\mathrm{y} - \mathrm{X}\beta) + \tau\sum_{j=1}^{k}|\beta_j|. \quad (19)$$

There were, however, ways to estimate the lasso coefficients. Indeed, Tibshirani provided an algorithm that finds the lasso solutions by treating the problem as a least squares problem with 2^k inequality constraints, and applying the constraints sequentially [17]. An even more attractive way to solve the lasso problem is proposed by Efron *et al.* [19], who solve the lasso problem with a small modification to the least angle regression (LARS) algorithm, which is a variation of the classic forward selection algorithm in linear regression. The modification ensures that the sign of any non-zero estimated regression coefficient is the same as the sign of the correlation coefficient between the corresponding explanatory variable and the current residuals. Grandvalet [20] showed that the lasso was equivalent to adaptive ridge regression and develops an expectation–maximization (EM) algorithm to compute the lasso solution. In some of the lasso algorithms, such as the modified LARS algorithm and the algorithm Tibshirani [17] described the shrinkage parameter δ (or z) must be estimated before finding the lasso solutions. Hastie *et al.* [21] estimated the parameter

$$\delta = \frac{\sum_{j=1}^{k}|\hat{\beta}_j|}{z} \quad (20)$$

through ten-fold cross-validation, where z was some positive scalar that reduced the ordinary least squares coefficient estimates. Tibshirani [17] used five-fold

cross-validation, generalized cross-validation, and a risk minimizer to estimate the parameter z, with the computational cost of the three methods decreasing in the same order. Efron *et al.* [19] also recommended using cross-validation to estimate the lasso parameter. If z was one or less, there was no shrinkage and the lasso solutions for the coefficients were the least squares solutions. One can also view the lasso shrinkage parameter as the fraction of the ordinary least squares solution that was the lasso solution. The lasso can be implemented in SAS relatively easily using GLMSELECT procedure [11].

4. COMPARISON OF METHODS ON SIMULATED DATA

4.1. Data Simulation and Analysis

To compare the relative performance of methods for dealing with collinearity (R-R, PC-R, PLS-R, L-R), we simulated datasets with various range of predictor collinearity. For our simulation experiment, we created training and test datasets that had sample size n (n = 1000) and 15 explanatory normally distributed variables (predictors).

These were grouped into three clusters (A, B, C). The cluster A involved five correlated variables (X1-X5), the cluster B included five variables (X6-X10) obtained by various combinations of variables of the previous cluster and at last, the cluster C that involved five (X11-X15) uncorrelated variables. In order to assess the performance of selected methods according to the degree of collinearity, we have simulated variables within of cluster A with very low (0.09) correlation, moderate (0.50), high (0.75) or very high (0.99) correlation. Then, to explore the performance of R-R, PC-R, PLS-R, L-R methods for dealing with collinearity according to its shape of the combinations of variables, we have generated variables within of cluster B as linear, exponential, quadratic or reciprocal combinations. We have also simulated variables for some cases more complex with both different levels of correlation and with various combinations as that may be often the case in practice. For all training and test data sets, the response variable was calculated as a function of predictors X1-X15 plus random normal noise (standard deviation = 0.1).

We analysed each data set with the selected collinearity methods (R-R, PC-R, PLS-R, L-R) and two other naïve or without correction methods, in which the analyses are performed ignoring exposures collinearity. The first of these two naïve methods is a standard multiple linear regression and the second method is a stepwise linear multiple regression. For each set, the estimates $\hat{\beta}$, the standard deviation of the estimates $\hat{\beta}$, as Root Mean Squared Error ($\mathbf{RMSE} = \sqrt{\frac{\sum (y - \hat{y})^2}{n}}$) and total MSE were derived.

4.2. Simulation Results

4.2.1. Performance of Correction Methods According to the Degree of Collinearity

Results of the analyses to assess the performance of selected methods (R-R, PC-R, PLS-R, L-R) according to the degree of collinearity are presented in Table **1**. Firstly, when the collinearity in exposures was ignored, as expected the stepwise linear multiple regression appears to be the best model for prediction compared to the standard multiple linear regression. It is also noted that the number of selected variables within of cluster A after the stepwise procedure varied according to the degree of collinearity. Indeed, when the correlations of variables within of cluster A were low (0.20) to high (0.70) all variables X1-X5 were considered after the stepwise procedure; that was not the case when the correlation was very high. Furthermore, it is also noted that regression coefficients were biased according to the degree of collinearity. Secondly, when the statistical approaches (R-R, PC-R, PLS-R, L-R) were used to address different levels of collinearity, we observed that R-R, PC-R and PLS-R methods appeared to have a somewhat similar behavior, with a slight advantage for the R-R and PLS-R methods. Indeed, in all implemented cases, the R-R and PLS-R methods always provided the smallest value of RMSE or AIC. As expected, regression coefficients were very close when the degree of collinearity is high, moderate or very low.

4.2.2. Performance of Correction Methods in Case of High Correlation and Various Shape of Collinearity

Results of the analyses to assess the performance of selected methods (R-R, PC-R, PLS-R, L-R) in case of high correlation and various shape of collinearity are presented in Table **2**. Also like in the previous analysis, it is also noted that regression coefficients are biased according to the degree and the shape of collinearity. The use of statistical approaches (R-R, PC-R, PLS-R, L-R) to address the collinearity problems appeared to have a somewhat similar performance, with a slight advantage for the PC-R and R-R methods.

Table 1: Performance of Correction Methods According to the Degree of Collinearity

		Without correction		After correction of collinearity			
		Naive analyses	**Naive analyses with stepwise**	**Ridge Regression**	**PC Regression**	**PLS Regression**	**Lasso Regression**
Very low correlation X1-X5 (0.09)	X1(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X2(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X3(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X4(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X5(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	RMSE	0.319	0.319	0.322	0.323	0.320	0.319
	AIC	-1275.5	-1275.5	-1254.0	-1262.1	-1275.4	-1275.5
	R^2 adjusted	0.47	0.47	0.46	0.46	0.47	0.47
Moderate correlation X1-X5 (0.50)	X1(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X2(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X3(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X4(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X5(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	RMSE	0.320	0.320	0.323	0.321	0.320	0.320
	AIC	-1275.5	-1275.5	-1253.7	-1270.0	-1274.5	-1275.5
	R^2 adjusted	0.62	0.62	0.61	0.62	0.62	0.62
High correlation X1-X5 (0.75)	X1(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X2(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X3(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X4(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X5(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	RMSE	0.320	0.320	0.325	0.321	0.320	0.320
	AIC	-1275.5	-1275.5	-1244.5	-1271.6	-1273.5	-1275.5
	R^2 adjusted	0.68	0.68	0.67	0.68	0.68	0.68
Very high correlation X1-X5 (0.99)	X1(β=0.1)	1.5	1.6	0.1	0.1	0.1	0.5
	X2(β=0.1)	-0.1	-	0.1	0.1	0.1	-
	X3(β=0.1)	0.6	-	0.1	0.1	0.1	-
	X4(β=0.1)	-1.2	-1.1	0.1	0.1	0.1	-
	X5(β=0.1)	-0.4	-	0.1	0.1	0.1	-
	RMSE	0.320	0.320	0.326	0.321	0.321	0.320
	AIC	-1275.5	-1275.0	-1235.4	-1271.5	-1271.4	-1272.7
	R^2 adjusted	0.73	0.73	0.72	0.73	0.73	0.73

Table 2: Performance of Correction Methods in Case of Complex Different Levels and Various Shape of Collinearity

		Without correction		After correction of collinearity			
		Naive analyses	Naive analyses with stepwise	Ridge Regression	PC Regression	PLS Regression	Lasso Regression
Very high correlation X1-X5 (0.99) and linear combinations X6-X10	X1(β=0.1)	-23.1	-	0.5	-0.2	-0.2	-
	X2(β=0.1)	-0.8	-1.6	0.4	-0.2	-0.3	-1.8
	X3(β=0.1)	-0.5	-	0.4	-0.2	-0.3	-1.4
	X4(β=0.1)	-0.2	-	0.4	-0.3	-0.3	-
	X5(β=0.1)	-0.4	-	0.4	-0.3	-0.3	0.5
	X6(β=0.1)	0.0	-	0.5	-0.2	-0.2	-
	X7(β=0.1)	0.0	-1.3	0.1	-0.1	-0.1	-
	X8(β=0.1)	0.0	-	0.0	0.0	0.0	-
	X9(β=0.1)	1.2	0.3	0.0	0.1	0.1	0.1
	X10(β=0.1)	0.0	0.9	0.3	1.1	1.1	1.1
	RMSE	0.179	0.179	0.317	0.182	0.182	0.179
	AIC	-2435.4	-2434.5	-1290.5	-2403.2	-2407.4	-2435.0
	R^2 adjusted	1.0	1.0	1.0	1.0	1.0	1.0
Very high correlation X1-X5 (0.99) and quadratic combinations X6-X10	X1(β=0.1)	1.8	2	0.3	0.8	0.9	-
	X2(β=0.1)	-4.6	-4.4	-1.4	-3.9	-4.2	-
	X3(β=0.1)	1.1	1.3	0.7	0.9	0.7	0.2
	X4(β=0.1)	-0.1	-	-0.1	-0.5	-0.1	-
	X5(β=0.1)	0.6	-	0.1	0.8	0.7	-
	X6(β=0.1)	0.0	-	0.3	0.8	0.9	-
	X7(β=0.1)	0.5	0.5	0.3	0.5	0.5	0.5
	X8(β=0.1)	-0.4	-0.4	-0.1	-0.4	-0.4	-
	X9(β=0.1)	0.3	0.3	0.3	0.3	0.3	0.1
	X10(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	RMSE	0.182	0.182	0.215	0.182	0.182	0.211
	AIC	-2400.2	-2399.6	-2067.7	-2398.8	-2400.1	-2107.3
	R^2 adjusted	1.0	1.0	1.0	1.0	1.0	1.0
Very high correlation X1-X5 (0.99) and reciprocal combinations X6-X10	X1(β=0.1)	0.3	-	0.1	0.1	0.1	0.2
	X2(β=0.1)	0.6	-	0.1	0.1	0.1	0.4
	X3(β=0.1)	0.1	-	0.1	0.1	0.1	-
	X4(β=0.1)	0.0	0.6	0.1	0.1	0.1	-
	X5(β=0.1)	-0.4	-	0.1	0.1	0.1	-
	X6(β=0.1)	0.0	-	0.1	0.1	0.1	-
	X7(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X8(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X9(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	X10(β=0.1)	0.1	0.1	0.1	0.1	0.1	0.1
	RMSE	0.371	0.371	0.372	0.371	0.371	0.371
	AIC	-978.1	-977.4	-976.2	-977.7	-977.7	-977.9
	R^2 adjusted	0.91	0.91	0.91	0.91	0.91	0.91

Table 3: Anthropometric Data of Study Populations

	N	Age (years) Mean ± σ [min - max]	Weight (kg) Mean ± σ [min - max]	Height (cm) Mean ± σ [min - max]	BSA (m^2) Mean ± σ [min - max]	BMI (kg/m^2) Mean ± σ [min - max]	Volume (cm^3) Mean ± σ [min - max]
Heart	270	30.2 ± 17.5 [0.7 - 82.9]	59.0 ± 20.5 [9.0 - 109.0]	161.0 ± 21.0 [60.0 - 197.0]	1.6 ± 0.4 [0.4 - 2.3]	21.9 ± 5.0 [12.4 - 43.6]	544.2 ± 184.1 [94.9 - 993.0]
Thyroid	187	36.8 ± 21.5 [1.7 - 88.8]	58.9 ± 21.8 [10.5 - 106.0]	159.5 ± 22.5 [82.0 - 197.0]	1.6 ± 0.4 [0.5 - 2.3]	22.1 ± 5.2 [12.3 - 38.3]	16.8 ± 9.4 [2.1 - 49.4]

5. APPLICATION TO REAL EXAMPLES

To illustrate the effects of collinearity on model selection across methods, we ran two case studies with real data in radiation epidemiology area. To assess the health risks of exposure to ionizing radiation, radiation dosimetry is required. In the case of patients treated by external beam radiation therapy (EBRT), the radiation dosimetry is of high importance because it helps to determine the steepness of the dose response curve both for organs in target volumes and normal tissue of organs apart from the target volume. In the present clinical practice, the use of computed tomography (CT) images in EBRT planning has allowed to determine the organs 3D volume using approaches based on contouring and localization of structures [22]. However, for scientific purposes, it may be requested to estimate the organs 3D volume although they do not justify to be including in the RT planning CT. Similarly, for patients treated early years, CT images in EBRT planning are not available, thus the organ volume is determined by another approach, which is often established based on clinical and anthropometric data patient's or anatomy phantoms references recommended by the International Commission on Radiological Protection (ICRP) [23]. The anthropometric data including body height, body weight, body mass index (BMI) and body surface area (BSA) that are usually available, are often used to predict the organ volume [24]. Details of the data used in this study are given in Table **3**. In all datasets attention was restricted to persons exposed in a childhood.

5.1. Datasets for Modelling Heart and Thyroid Volumes

The Dataset for modelling heart volume was the one used in the analysis by Badouna *et al.* [22]. In this dataset, 270 patients treated for a childhood cancer at IGR, Villejuif, France, between 2003 and 2010, with RT planning data including a CT of the thorax archived in the Picture Archive and Communication System. CT data were acquired before EBRT during the treatment planning procedure. These thoracic scanners were used to determine the heart volume of these patients. Further details on the dataset, image segmentation and total heart volume reconstruction are given in the paper of Badouna *et al.* [22]. The Dataset for modelling Thyroid volume was that used in the analysis by Veres *et al.* [25]. In this dataset, 187 patients treated for a childhood cancer at IGR, with RT planning data including a CT of the neck archived in the Picture Archive and Communication System. CT data were acquired also before EBRT during the treatment planning procedure. These neck scanners were used to determine the heart volume of these patients. Further details on the dataset, image segmentation and volume of the thyroid reconstruction are given in the paper of Veres *et al.* [25].

5.2. Approaches for Collinearity in Heart and Thyroid Volumes Modelling

The correlations matrixes for the heart and thyroid database are provided in Table **4**. This table showed, among other things, that in the heart database, the correlation between the weight and height was high. As expected BSA was very strongly correlated to both weight and height, while BMI was very strongly correlated to weight alone and had moderate correlation with height. Similar results were obtained in the Thyroid database. Hence, in this situation of collinearity that violated the assumption of independence of predictors, the parameter estimates derived from heart and thyroid volumes modelling may be unstable and untrustworthy.

Table **5** presents naïve and corrected results of the analyses of heart and thyroid volumes modelling. As in the previous naïve analysis, the stepwise linear multiple regression appears to be the best model for prediction compared to the standard multiple linear regression (in terms of RMSE and R^2 adjusted). When the statistical approaches (R-R, PC-R, PLS-R, L-R)

Table 4: Correlation Matrix

		Age	Weight	Height	BSA	BMI
Heart database	Age	1				
	Weight	0.5489 <.0001	1			
	Height	0.4698 <.0001	0.8116 <.0001	1		
	BSA	0.5585 <.0001	0.9826 <.0001	0.9029 <.0001	1	
	BMI	0.5004 <.0001	0.8541 <.0001	0.4238 <.0001	0.7632 <.0001	1
Thyroid database	Age	1				
	Weight	0.6233 <.0001	1			
	Height	0.5607 <.0001	0.8233 <.0001	1		
	BSA	0.6387 <.0001	0.9837 <.0001	0.9092 <.0001	1	
	BMI	0.5366 <.0001	0.8671 <.0001	0.4681 <.0001	0.7848 <.0001	1

Table 5: Correction of Collinearity in Heart and Thyroid Volumes Modelling

		Without correction		After correction of collinearity			
		Naive analyses β(SE)	**Naive analyses with stepwise β(SE)**	**Ridge Regression β(SE)**	**PC Regression β(SE)**	**PLS Regression β(SE)**	**Lasso Regression β(SE)**
Heart database	Intercept	3.913(0.32)*	3.821(0.15)*	3.774(0.13)*	3.084(0.13)*	3.105(0.13)*	3.842(0.15)*
	Age	0.003(0.00)*	0.003(0.00)*	0.004(0.00)*	0.004(0.00)*	0.003(0.00)*	0.003(0.00)*
	Gender	-0.103(0.02)*	-0.103(0.02)*	-0.080(0.02)*	-0.098(0.02)*	-0.097(0.02)*	-0.103(0.02)*
	Weight	-0.032(0.01)*	-0.030(0.00)*	-0.001(0.00)*	-0.006(0.00)*	-0.009(0.00)*	-0.030(0.00)*
	Height	-0.156(0.49)	-	1.037(0.07)*	1.564(0.07)*	1.450(0.07)*	-
	BSA	2.597(0.64)*	2.401(0.20)*	0.336(0.03)*	0.184(0.03)*	0.390(0.03)*	2.373(0.19)*
	BMI	0.010(0.01)	0.012(0.01)*	0.009(0.00)*	0.030(0.00)*	0.029(0.00)*	0.011(0.01)*
	R^2 adjusted	**0.909**	**0.910**	**0.909**	**0.910**	**0.907**	**0.910**
	RMSE	**0.129**	**0.128**	**0.132**	**0.128**	**0.130**	**0.128**
Thyroid database	Intercept	0.231(1.00)	-0.335(0.23)	-0.110 (0.25)	-0.097(0.25)	-0.315(0.25)	-0.284(0.21)
	Age	0.003(0.00)	0.003(0.00)*	0.003(0.00)*	0.003(0.00)*	0.003(0.00)*	0.003(0.00)*
	Gender	-0.047(0.05)	-	-0.042(0.05)	-0.038(0.05)	-0.045(0.05)	-0.045(0.05)
	Weight	-0.013(0.02)	-	0.004(0.00)*	0.005(0.00)*	0.002(0.00)*	-
	Height	-0.193(1.6)	1.191(0.26)*	1.141(0.17)*	1.152(0.17)*	1.271(0.17)*	1.167(0.26)*
	BSA	1.922(2.21)	0.597(0.15)*	0.409(0.05)*	0.398(0.05)*	0.440(0.05)*	0.603(0.16)*
	BMI	-0.014(0.03)	-0.014(0.03)	-0.004(0.01)*	-0.007(0.01)*	0.000(0.01)*	-
	R^2 adjusted	**0.727**	**0.730**	**0.734**	**0.729**	**0.734**	**0.730**
	RMSE	**0.333**	**0.332**	**0.329**	**0.332**	**0.329**	**0.332**

*β statistically significantly at the 5% level.

were used to address the collinearity, we observed that results from naïve and corrected methods were quite similar, but the use of correction methods can overcome the violation of the assumption of independence of predictors. When investigating in detail the performance of selected corrected methods, we found that it was somewhat similar in terms of RMSE and R^2 adjusted, with a slight advantage for the PC-R and Lasso-R methods from the heart database and with a slight advantage for the Ridge-R and PLS-R methods from the Thyroid database.

6. DISCUSSION

6.1. Heart and Thyroid Volumes Modelling

Previously, studies of Veres *et al.* [25] and Badouna *et al.* [22] provided evidence of a relationship between thyroid or heart volume and anthropometric measurements (weight, height, BMI and BSA), gender and age. These authors have also developed some prediction equations models using anthropometric measurements thyroid or heart volume in human models used to represent external beam radiotherapy (EBRT) patients. Indeed, for patients treated early years, for whom CT images in EBRT planning were not available, a modelling approach of organ volume based on clinical and anthropometric data patient's may be better that recourse to the anatomy phantoms references. Because, the volume of some organs may be subject to inter-individual variations which must be taken into account when evaluating doses [26, 27]. In addition, Scarboro *et al.* demonstrated that better knowledge on organ volume could potentially impact the design of epidemiological studies of a radiation-induced late effect for organs that are known to vary in size between individuals [28]. Many authors have shown that anthropometric measurements (weight, height, BMI or BSA) were significantly associated to some organs volume. For instance, in adults, the volume of the thyroid gland significantly increases not only with weight and height, but also with BMI and BSA [29]. According to Veres *et al.*, the best fit for children was obtained by modelling the log of thyroid volume as a linear function of body surface area (BSA) and age and for adults, as a linear function of BSA and gender [25]. However, Badouna *et al.* reported that, among anthropometric parameters, weight plays an important role in predicting heart volume [22]. In these previous studies, the issue of multicollinearity was not dealt. Therefore, the parameter estimates derived from heart and thyroid volumes modelling may be unstable and untrustworthy. Nevertheless, the use of variable selection procedures to select the "best" subset of variables, may allow indirectly reducing the effects of multicollinearity in the final model.

6.2. Methods to Address Collinearity in Statistical Models and Practical Recommendations

Collinearity is a problem recognised by most introductory textbooks on statistics [4]. A broad variety of methods to address collinearity problems have been developed, but, despite the relevance of the problem, these methods have rarely been applied in everyday routine. The Ridge regression (R-R) method is the first method used to deal with collinearity in covariates, mainly because of its relative simplicity compared to others methods. There is a large body of literature that illustrates the multifaceted of this method [30]. Ridge regression differs in two aspects from other techniques. It does not extract orthogonal components called latent variables or latent factors and it applies explicit shrinkage to the regression vector. It has been shown that it can compete quite well with PLS-R and PC-R methods with regard to prediction performance [31]. Our results are on the whole consistent with this observation. Using simulation studies of complex chemical mixtures to compare PLS-R and PC-R methods, Wentzell and Montoto reported no significant differences in prediction errors between these methods [32]. As some authors, our study showed that there were fewer numbers of latent variables retained with PLS-R than PC-R. With some misuse of language, we can say that PLS-R method was clearly more parsimonious than PC-R. On the other hand, PLS-R can be preferred over other correction methods because this method can accept multiple response variables. If data both for modelling heart and thyroid volumes were based on the same subjects, the PLS-R method will remain the only one that is most suitable. Moreover, like stepwise regression, the lasso (L-R) can explore models with more covariates than observations, but the lasso is a "less greedy" procedure than stepwise regression in that it tends to find less complex models [32]. Tibshirani [17] argued that if there are multicollinearity among predictors, R-R dominates the lasso (L-R) in prediction performance. This statement was somewhat mitigated in our study because, we found that the two methods are not per se different in terms of prediction, with in some cases better performance for L-R.

Our work did not consider some multicollinearity-correction methods, for instance the Bayesian regression methods that could be an ideal solution for

this problem according to Xu [34]. In addition, the Bayesian method can be very efficient when p >> N [35]. However, the lack of statistical standard software to implement Bayesian regression methods to deal with multicollinearity and the convergence problems encountered often with Bayesian, ruled out the consideration of Bayesian methods in our study. Moreover, further investigation regarding the four methods (R-R, PC-R, PLS-R, L-R) considered in the present work and other methods like the elastic net, the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) and Bayesian method might be worthwhile for other models than the linear [35].

7. CONCLUSION

The aim of this paper was to compare the behavior of four main techniques to address multicollinearity (R-R, PC-R, PLS-R, L-R). Comparisons were performed using a simulation study and within two datasets used for modelling organs volume as a function of clinical and anthropometric parameters. Our simulation results showed that regression coefficients were biased according to the degree of collinearity, in particular under severe collinearity. Similarly, when the collinearity structure changed non-linearly or was completely lost, however, model fit decreased substantially. Overall, the RMSE was slightly lower and the R^2 adjusted was slightly higher after correction than the naïve results. The performance of four correction methods implemented was very close, but with a slight advantage for PC-R and PLS-R methods in most scenarios implemented. Furthermore, the application of correction methods to the heart and Thyroid databases allowed us to provide stable and trustworthy parameter estimates for heart and thyroid volumes predictors. Recommendations to deal with the problem of multicollinearity in epidemiological studies have been expressed many years ago [36]. In practice, however, despite the ubiquity of multicollinearity, the use of methods for multicollinearity correction is still scarce. One important reason is that despite many proposed methods, little is known about their strength or performance. Therefore, this work will contribute to highlighting performances of methods used only for situations ranging from low to very high multicollinearity.

ACKNOWLEDGEMENTS

We apologize to authors whose relevant publications were not cited due to space limitation. This work was supported by the French Agence Nationale Pour la Recherche Scientifique (Hope-Epi project), ARC foundation with the Pop-HaRC project, INCa/ARC foundation with the CHART project, European Commission, FP7-Health, PanCaresurf-Up project, INSERM Plan Cancer PeriDoseQuality project, Foundation Pfizer for childhood and adolescent health, Ligue Nationale Contre le Cancer, the Institut de Recherche en Santé Publique, Programme Hospitalier de Recherche Clinique, Agence Française de Sécurité Sanitaire et Produit de Santé. These funding agencies had no role in the design and conduct of the study, in the collection, management, analysis and interpretation of the data, or in the preparation, review, and approval of the manuscript.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

REFERENCES

[1] Pitard A, Viel JF. Some methods to address collinearity among pollutants in epidemiological time series. Statistics in Medicine 1997; 16(5): 527-44. https://doi.org/10.1002/(SICI)1097-0258(19970315)16:5<527::AID-SIM429>3.0.CO;2-C

[2] Schroeder, Mary Ann. Diagnosing and dealing with multicollinearity. Western Journal of Nursing Research 1990; 12(2): 175-187. https://doi.org/10.1177/019394599001200204

[3] Gordon RA. Issues in multiple regression. American Journal of Sociology 1968; 73: 592-616. https://doi.org/10.1086/224533

[4] Dormann CF, Elith J, Bacher S, *et al*. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 2012; 35: 001-020.

[5] Weisberg S. Applied Linear Regression, third edition. New-York: Wiley. 2005. www.stat.umn.edu/alr

[6] Buonaccorsi JP. A modified estimating equation approach to correcting for measurement error in regression. Biometrika 1996; 83: 433-440. https://doi.org/10.1093/biomet/83.2.433

[7] Hoerl E, Kennard RW. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics 1970; 12: 69-82. https://doi.org/10.1080/00401706.1970.10488635

[8] Guilkey DK, Murphy JL. Directed Ridge Regression Techniques in cases of Multicollinearity. Journal of American Statistical Association 1975; 70: 767-775. https://doi.org/10.1080/01621459.1975.10480301

[9] El-Dereny M, Rashwan NI. Solving Multicollinearity Problem Using Ridge Regression Models. International Journal of Contemporary Mathematical Sciences 2011; 6: 585-600.

[10] Meijer RJ, Goeman JJ. Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. Biometrical Journal 2013; 55: 141-55. https://doi.org/10.1002/bimj.201200088

[11] SAS Institute Inc. SAS® 9.3 System Options: Reference, Second Edition. Cary, NC: SAS Institute Inc. 2011. https://support.sas.com/documentation/cdl/en/lesysoptsref/64892/PDF/default/lesysoptsref.pdf

[12] Vigneau E, Bertrand D, Qannari EM. Application of latent root regression for calibration in near-infrared spectroscopy: Comparion with principal component regression and partial

least squares. Chemometrics and Intelligent laboratory system 1996; 35: 231-238. https://doi.org/10.1016/S0169-7439(96)00051-2

[13] Cassel C, Westlund AH, Hackl P. Robustness of partial least-squares method for estimating latent variable quality structures. Journal of Applied Statistics 1999; 26: 435-448. https://doi.org/10.1080/02664769922322

[14] Chun H, Keleş S. Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. Journal of the Royal Statistical Society B Statistical Methodology 2010; 72: 3-25. https://doi.org/10.1111/j.1467-9868.2009.00723.x

[15] Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 2007; 8: 32-44. https://doi.org/10.1093/bib/bbl016

[16] Helland I. On the structure of Partial Least Squares. Communications in Statistics - Simulation and Computation 1988; 17: 581-607. https://doi.org/10.1080/03610918808812681

[17] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society B Statistical Methodology 2011; 73(3): 267-288. https://doi.org/10.1111/j.1467-9868.2011.00771.x

[18] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society B Statistical Methodology 2005; 67: 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[19] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Annals of Statistics 2004; 32: 407-451. https://doi.org/10.1214/009053604000000067

[20] Grandvalet Y. Least absolute shrinkage is equivalent to quadratic penalization. In Niklasson L, Boden M, Ziemske T (eds) ICANN'98 Perspectives in Neural Computing. Springer-Verlag: Berlin 1998. https://doi.org/10.1007/978-1-4471-1599-1_27

[21] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag. New York 2001. https://doi.org/10.1007/978-0-387-21606-5

[22] Badouna AN, Veres C, Haddy N, *et al*. Total heart volume as a function of clinical and anthropometric parameters in a population of external beam radiation therapy patients. Physics in Medicine & Biology 2012; 57: 473-484. https://doi.org/10.1088/0031-9155/57/2/473

[23] International Commission on Radiological Protection (ICRP). Basic Anatomical and Physiological Data for Use in Radiological Protection: Reference Values ICRP Publication 89 (Pergamon: Oxford) 2002.

[24] Graham TP Jr, Jarmakani JM, Canent RV Jr, *et al*. Left heart volume estimation in infancy and childhood. Reevaluation of methodology and normal values. Circulation 1971; 43: 895-904. https://doi.org/10.1161/01.CIR.43.6.895

[25] Veres C, Garsi JP, Rubino C, *et al*. Thyroid volume measurement in external beam radiotherapy patients using CT imaging: correlation with clinical and anthropometric characteristics. Physics in Medicine & Biology 2010; 55: 507-519. https://doi.org/10.1088/0031-9155/55/21/N02

[26] Xu XG, Bednarz B, Paganetti H. A review of dosimetry studies on external-beam radiation treatment with respect to second cancer induction. Physics in Medicine & Biology 2008; 53: 193-241. https://doi.org/10.1088/0031-9155/53/13/R01

[27] Zaidi H, Xu XG. Computational anthropomorphic models of the human anatomy: the path to realistic Monte Carlo modelling in radiological sciences. Annual Review of Biomedical Engineering 2007; 9: 471-500. https://doi.org/10.1146/annurev.bioeng.9.060906.151934

[28] Scarboro SB, Stovall M, White A, *et al*. Effect of organ size and position on out-of-field dose distributions during radiation therapy. Physics in Medicine & Biology 2010; 55: 7025-7036. https://doi.org/10.1088/0031-9155/55/23/S05

[29] Barrère X, Valeix P, Preziosi P, Bensimon M, Pelletier B, Galan P, Hercberg S. Determinants of thyroid volume in healthy French adults participating in the SU.VI.MAX cohort. Clinical Endocrinology 2000; 52: 273-278. https://doi.org/10.1046/j.1365-2265.2000.00939.x

[30] Gruber MHJ. Regression Estimators: a Comparative Study. Academic Press: Boston 1990.

[31] Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics 1993; 35: 109-148. https://doi.org/10.1080/00401706.1993.10485033

[32] Wentzell PD, Montoto V. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. Chemometrics and Intelligent Laboratory Systems 2003; 65: 257-279. https://doi.org/10.1016/S0169-7439(02)00138-7

[33] Wu J, Devlin B, Ringquist S, Trucco M, Roeder K. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic Epidemiology 2010; 34: 275-85. https://doi.org/10.1002/gepi.20459

[34] Xu S. Estimating polygenic effects using markers of the entire genome. Genetics 2003; 163: 789-801.

[35] Curtis SM, Ghosh SK. A Bayesian Approach to Multicollinearity and the Simultaneous Selection and Clustering of Predictors in Linear Regression. Journal of Statistical Theory and Practice 2011; 5: 715-735. https://doi.org/10.1080/15598608.2011.10483741

[36] Willis CE, Perlack RD. Multicollinearity: effects, symptoms, and remedies. Northeastern Journal of Agricultural and Resource Economics 1978; 7: 55-61. https://doi.org/10.1017/S0163548400001989

Received on 15-04-2018 Accepted on 22-04-2018 Published on 08-05-2018

https://doi.org/10.6000/1929-6029.2018.07.02.2