IN BRIEF

- Acquire an understanding of the concepts surrounding 'collinearity'.
- Appreciate the indications and symptoms of collinearity in multivariable regression.
- Become aware of the available diagnostic tools for collinearity.
- Gain knowledge in the assessment of collinearity in the dental literature.
- Learn of some solutions to overcome the problem of collinearity.

# Problems of correlations between explanatory variables in multiple regression analyses in the dental literature

Y-K. Tu,[1] M. Kellett,[2] V. Clerehugh[3] and M. S. Gilthorpe[4]

Multivariable analysis is a widely used statistical methodology for investigating associations amongst clinical variables. However, the problems of collinearity and multicollinearity, which can give rise to spurious results, have in the past frequently been disregarded in dental research. This article illustrates and explains the problems which may be encountered, in the hope of increasing awareness and understanding of these issues, thereby improving the quality of the statistical analyses undertaken in dental research. Three examples from different clinical dental specialities are used to demonstrate how to diagnose the problem of collinearity/multicollinearity in multiple regression analyses and to illustrate how collinearity/multicollinearity can seriously distort the model development process. Lack of awareness of these problems can give rise to misleading results and erroneous interpretations. Multivariable analysis is a useful tool for dental research, though only if its users thoroughly understand the assumptions and limitations of these methods. It would benefit evidence-based dentistry enormously if researchers were more aware of both the complexities involved in multiple regression when using these methods and of the need for expert statistical consultation in developing study design and selecting appropriate statistical methodologies.

## INTRODUCTION

Multivariable statistical methods, such as multiple linear or logistic regression, have become widely used to analyse data in dental research. However, reduction in the effort required to complete calculations, due to the power of modern computers, does not imply that the required understanding of the statistical methods and assumptions that underpin regression analyses are similarly reduced. Medical statisticians have repeatedly warned against the misuses of correlation and regression within medical and dental research;[1,2] sometimes correlation and simple regression may give rise to spurious results if researchers do not comprehend fully the underlying statistical theory.[3-5]

One common problem in the use of multiple linear or logistic regression when analysing clinical data is the occurrence of explanatory variables (covariates) which are not independent, ie correlations amongst covariates are not zero.[6] Most textbooks emphasise that there should be no significant associations between covariates, as this gives rise to the problem known as *collinearity*.[6-11] When there are more than two covariates that are highly correlated, this is *multicollinearity*. Collinearity and multicollinearity can seriously distort the interpretation of a model. The role of each covariate is to cause increased inaccuracy, as expressed through bias within the regression coefficients,[11] and increased uncertainty, as expressed through coefficient standard errors.[6,7] Consequently, regression coefficients biased by collinearity 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

[1*]Clinical Reesearch Fellow, Department of Periodontology, Division of Restorative Dentistry, Leeds Dental Institute, University of Leeds, Leeds LS2 9LU and Biostatics Unit, Centre for Epidemiology and Biostatistics, University of Leeds, Leeds LS2 9LN; [2]Dean/Director of Leeds Dental Institute and Consultant in Restorative Dentistry, Division of Restorative Dentistry, Leeds Dental Institute, University of Leeds; [3]Professor of Periodontology, Department of Periodontology, Division of Restorative Dentistry, Leeds Dental Institute, University of Leeds; [4]Reader in Statistical Epidemiology, Biostatics Unit, Centre for Epidemiology and Biostatics, University of Leeds.
*Correspondence to: Dr Yu-Kang Tu
Email: y.k.tu@leeds.ac.uk.

**Refereed Paper**
**Accepted 9 May 05**
**doi: 10.1038/sj.bdj.4812743**

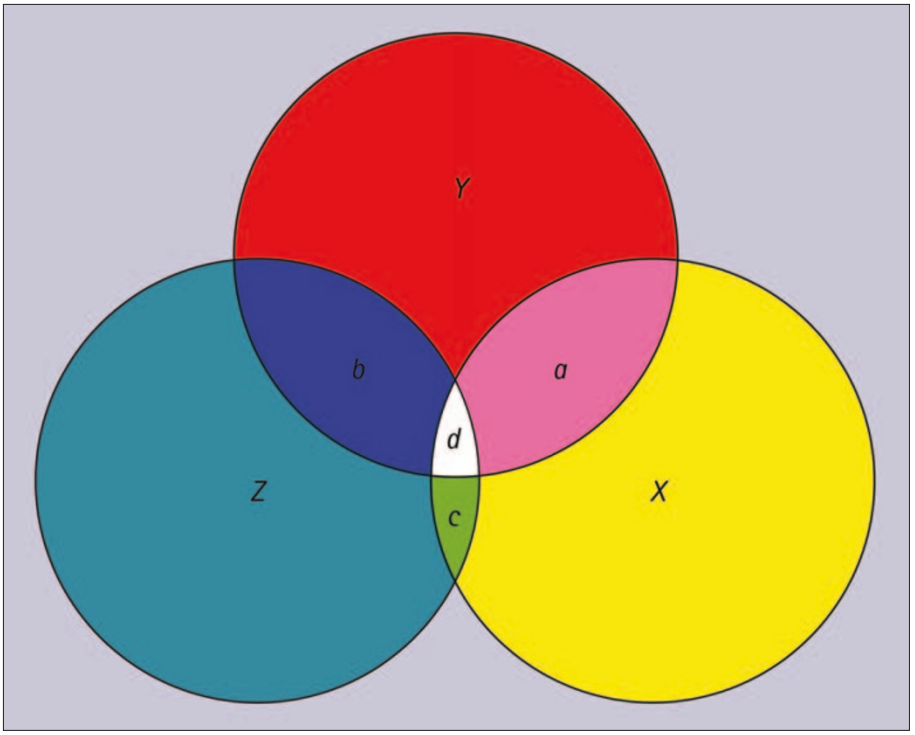


**Fig. 1a Venn's Diagram for the scenario where the correlation between covariates X and Z is small**

false-negative results (Type II error).

For instance, classical examples used by many textbooks to illustrate multicollinearity are where several explanatory variables are significantly correlated with the outcome variable using correlation or simple regression. Within a multiple regression model, none or few of the covariates are statistically significant, yet the overall variance of the dependent variable explained by the covariates is high (this is measured by $R^2$). This is because the information given by each covariate 'overlaps' with other covariates, due to multicollinearity.[12] Thus, it becomes hard, if not impossible, to distinguish amongst the individual contributions of each covariate to the outcome variance.

It is helpful to use Venn's diagrams to illustrate the problems of collinearity in a regression model where Y is regressed on X and Z (Figs 1a and 1b). Each circle is the variance of the variables. The overlapped area is the covariance between the two or three variables. For instance, b + d is the covariance between Y and Z. Multiple regression seeks to estimate the **independent** contribution of X and Z to the variance of Y, ie to estimate a and b for X and Z, respectively. Figure 1a shows the scenario where the correlation between X and Z is small, ie c and d are relatively small compared to a and b. Figure 1b shows the scenario where the correlation between X and Z is high, ie c and d are large. Although the correlations between Y and X and between Y and Z remain similar and the total explained variance (a, b, and d) of Y by X and Z remain similar, a large correlation between X and Z makes a and b become smaller and statistically non-significant.

However, an important point often overlooked is that even when regression coefficients are statistically significant, collinearity and multicollinearity can cause serious problems in the interpretation of results from a regression analysis. For instance, the relationship between the outcome and a covariate might be reversed when another covariate is entertained into the model.

The aim of this article is to provide a non-technical introduction to the concepts of collinearity and multicollinearity, and use several examples from the dental literature to demonstrate how to diagnose the problems of collinearity and multicollinearity in regression analysis. This article can be considered as an addition to the series of *further statistics* in dentistry in this journal.[13,14] For readers with mathematical minds, technical explanations of the problems of collinearity and multicollinearity can be found in our previous article[15] and advanced statistical textbooks.[7,8,10,11,16]

## Collinearity

Consider a multiple regression model with two covariates:

$$y=b_0+b_1x_1+b_2x_2;$$

where $y$ is the outcome variable (also known as the dependent variable), $x_1$ and $x_2$ are two covariates (also known as explanatory variables or independent variables), $b_0$ is the intercept, and $b_1$ and $b_2$ are regression coefficients for $x_1$ and $x_2$ respectively. Ideally, the best model for $y$ is that the correlation between $x_1$ and $x_2$ is zero, yet both $x_1$ and $x_2$ are highly correlated with $y$. If $x_1$ and $x_2$ are highly correlated with each other, and the direction of their correlation is the same as their respective correlations with $y$, collinearity might be a problem. This is because most of the contribution of $x_1$ and $x_2$ in explaining variation in the outcome, or in predicting $y$, 'overlaps'. Then it becomes difficult to quantify the individual contribution of $x_1$ and $x_2$, which is expressed through reduced regression coefficients and inflated standard errors.[6] When the correlation between $x_1$ and $x_2$ is exactly one, the situation is called *perfect collinearity*, and one covariate needs to be removed from the regression model in order to estimate a solution.[8] As $R^2$ can always be increased by adding a covariate to a linear regression (this is why an adjusted $R^2$ is also given in the regression output), $R^2$ can be large (ie close to one, as $R^2$ cannot exceed one) when there is serious multicollinearity in a model with many covariates, despite few covariates demonstrating statistical significance.[7-9]

## Multicollinearity

In a multiple regression model with $k$ covariates ($k>2$), ie: $y=b_0+b_1x_1+b_2x_2+...+b_kx_k$, the problem of multicollinearity is more complex and more difficult to detect, because multicollinearity does not necessarily require high bivariate correlations between covariates. For instance, if $x_1$, $x_2$ and $x_3$ are independent, the bivariate correlations between each pair are zero. However, if a new variable $x_4$ is derived from $x_1$, $x_2$, and $x_3$, such that $x_4=x_1+x_2+x_3$, there is perfect multicollinearity amongst the four variables, since each $x_i$ ($i$=1 to 4) can be expressed as a combination of the other three such as: $x_3=x_4-x_1-x_2$. Each pair of correlations between $x_4$ and the remaining three covariates may be relatively modest, but multicollinearity is still a serious problem due to the fact that the information provided by the four variables as a whole is overlapped. Unless one of the four covariates is removed from the regression model, computer software cannot proceed with mathematical computation for the regression model.

## Diagnosis of multicollinearity

One of the diagnostic methods for multicollinearity is to perform auxiliary regressions, to regress one covariate on the remaining covariates.[7,8] For instance, $x_4$ is used as the outcome and is regressed on $x_1$, $x_2$, and $x_3$, the $R^2$ for this auxiliary regression is a measure of the degree of multicollinearity for $x_4$. The variance inflation factor (*VIF*), defined as $VIF=1/(1-R^2_i)$ where $R^2_i$ is the $R^2$ for a covariate $x_i$ regressed on the remaining covariates in a auxiliary regression, is the most commonly used

regression diagnostic for multicollinearity within standard statistical software.[7-11]

Another diagnostic tool for multicollinearity is the *condition index*, which is more complicated but provides very similar information as the *VIF*.[8,14] Detailed explanations of these diagnostics can be found in the cited references.[7-11] In general, standard errors of regression coefficients are inflated when the *VIF* is large (eg when *VIF*>10, multicollinearity is usually considered a problem, though this is an arbitrary threshold).

In summary, when there are more than two covariates in a regression model, correlations amongst covariates are informative but should not be the only criterion used to judge whether or not multicollinearity is a problem. Other diagnostic tools, such as the *VIF* and *condition index* should also be used and reported. Moreover, even when there is a problem of multicollinearity, the collinear covariates may remain statistically significant, though the sign of regression coefficient might be contrary to expectation – this is another indication of potential problems due to multicollinearity.[7-10]

In the next sections, we use three examples within dental research to illustrate how to detect the problem of collinearity. It should be noted that these examples were selected as they exhibit good quality in the reporting of their regression analyses, thereby providing sufficient information to the reader to evaluate whether or not collinearity is a potential problem. This is frequently not the case in current clinical research publications.

### Example one: *Mutans Streptococci* in plaque and saliva

In a study to investigate the association between caries incidence and *Mutans Streptococci* (MS) scores,[17] simple logistic regression indicated increased odds of experiencing new caries amongst children with higher plaque MS scores (Odds Ratio (OR) = 15.26, 95% Confidence Interval (CI) = 6.52, 38.78) and higher saliva MS scores (OR = 5.78, 95% CI = 2.66, 13.12) than children with lower plaque and saliva MS scores; ie children with higher baseline plaque or saliva MS scores had greater experience of new caries when re-examined six months later. However, in a multiple logistic regression model, the plaque MS score OR = 12.59 (95% CI = 3.18, 67.08) whilst the saliva MS score OR = 0.48 (95% CI = 0.09, 1.95). Since OR<1 signals a reverse association between the levels of MS scores in saliva and the experience of new caries, this seems to suggest that, when adjusting for plaque MS scores (and other covariates), a high saliva MS score may have a small (though not statistically significant) protective effect on caries incidence.

Both plaque and saliva MS scores were positively associated with caries incidence in separate logistic regression models, or when evaluated using the chi-squared test. However, the multiple logistic regression model might suffer collinearity between the two MS scores, and the change in the direction of the association between caries incidence and saliva MS scores from the simple to the multiple logistic regression model might be nothing more than a symptom of collinearity. The statistical association can be verified by performing a chi-squared test for the two MS scores. Using the statistical package *R* (Version 2.0.0, *R* development core team, Vienna, Austria 2004) to perform a chi-squared test on Table 2 in the original article, the association between the two MS scores is highly significant ($\chi^2$ = 62.4 with one degree of freedom, $P$<0.0001). Therefore, it might be more appropriate to consider only plaque MS scores within the regression model.

This example indicates the problem in detecting collinearity between categorical variables. Since the Pearson moment-product correlation is only appropriate for continuous variables, many researchers overlook that collinearity and multicollinearity can arise when the association between categorical variables is strong. Appropriate statistical methods, such as the chi-squared test, should be used to detect the association between categorical variables.

### Example two: Number of missing teeth at baseline and subsequent tooth loss

Changes in the direction of association between the dependent variable and explanatory variable from simple to multiple regression is a common symptom of collinearity. In a prospective study to investigate the relationship between potential risk factors and subsequent tooth loss,[18] bivariate correlations showed that tooth loss over the period of 20 years (between 1970 and 1990) is positively correlated with: the marginal bone loss (MBL) index in 1970 ($r$=0.49; $P$<0.001); age in 1970 ($r$=0.21; $P$<0.001); Russell's index in 1970 ($r$=0.46; $P$<0.001); and the number of missing teeth in 1970 ($r$=0.08; $P$=0.038). However, stepwise multiple regression showed that regression coefficients for age (–0.039; $P$=0.021) and number of missing teeth (–0.094, $P$=0.003) were negative after adjusting for MBL index, Russell's index, and other baseline variables. It is apparent that the baseline risk factors measured in 1970 are highly correlated, as they are different manifestations of the same underlying (periodontal) disease in each patient. Therefore, interpretation of the unexpected negative associations from the multiple regression model needs to be made with extreme caution. This also indicates that, although associations between the outcome and the explanatory variables are reversed due to collinearity, $P$-values may still be small and hence highly significant.

### Example three: Horizontal bone fill and pocket depth, clinical attachment level and gingival margin position

In a study using guided tissue regeneration (GTR) to treat molar furcation defects,[19] multiple linear regression was performed to investigate the association between treatment outcome, horizontal bone fill, and six baseline measurements: pocket probing depth (PPD), clinical attachment level (CAL), gingival margin position (GMP), distance

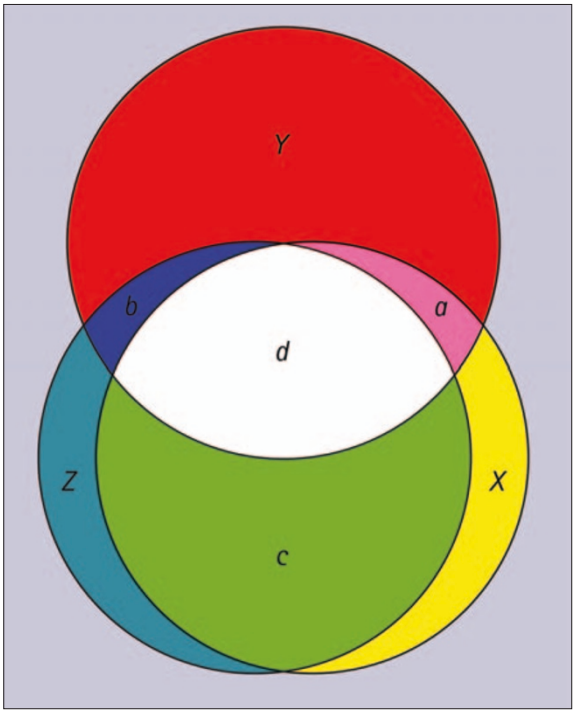


**Fig. 1b Venn's diagram for the scenario where the correlation between covariates X and Z is large**

between cemento-enamel junction to alveolar crest (CEJ-AC), vertical intrabony component (VIC), and horizontal defect depth (HDD). Results from the regression analysis revealed that treatment outcomes were significantly associated with baseline HDD in both treatment groups. As there is mathematical coupling[3-5] between baseline HDD and the outcome, horizontal bone fill (ie change in HDD), further statistical analyses are warranted to support this purported association. In general, mathematical coupling occurs when one variable directly or indirectly contains the whole or part of another, and the two variables are then analysed using correlation or regression,[4] such as investigating the relation between change or percentage change in variables (eg following an intervention) and their initial/baseline values (ie prior to the intervention).[3-5] Consequently, the statistical procedure of testing the null hypothesis – that the coefficient of correlation or the slope of regression is zero – becomes inappropriate.

However, notwithstanding mathematical coupling, Table 2 in the original article shows that in the models for each treatment group there was one covariate whose regression coefficient was absent. The original table used *NA*, which was probably the abbreviation for 'not available' or 'not applicable' (though no explanation was given in the original article regarding why these regression coefficients were not available). This illustrates how perfect multicollinearity is frequently overlooked because most statistical software will (if required to proceed automatically) remove one of the perfectly collinear covariates in order to achieve meaningful model estimates of all remaining covariate coefficients. Some researchers perhaps fail to pay sufficient heed to the warnings that often accompany the regression output of many software packages when perfect multicollinearity is present. In this example, CAL = PPD + GMP, and therefore one of these three variables must be dropped from the model for estimation to proceed meaningfully. It is curious to note that, if executed with slightly different data within the same statistical software package, the final model might in fact exclude a different covariate: for the treatment group of GTR, CAL was removed, but for the treatment group of GTR combined with bone grafting, PPD was removed.

## SOLUTIONS

### Removal of redundant explanatory variables

The problems of collinearity and multicollinearity in the three examples might be diagnosed using either the *VIF* or the *condition index*. Although *VIF* > 10 is the criterion most often suggested by the textbooks, this is not, in our opinion, the only criterion to be used. The unexpected direction of associations between the outcome and explanatory variables is an important sign of collinearity and multicollinearity. When the direction of association differs between simple correlation/regression and multivariable regression, this does not necessarily indicate that the research has found intriguing results. On the contrary, researchers should carefully examine the relations between all the explanatory variables in the regression models. If some of the collinear variables are redundant, in terms of providing no extra useful information, or are simply duplicate measurements of the same variable, a solution is to remove these variables from the model. For instance, in periodontics, the assessment of extent of periodontal breakdown can be made clinically or radiographically, and these two measurements seem to be highly correlated. To include both variables in the same model probably does more harm than good from a statistical viewpoint.

### Centring

Multicollinearity can be a problem for a covariate when included in a model along with its quadratic form in a non-linear regression or when also included through a product-interaction term with another variable.[7-10] For instance, if the research question is whether or not the number of cigarettes smoked and the amount of alcohol consumed have a synergistic effect on the risk of oral cancer, a product term – *smoking-alcohol* – might be generated and entered as an additional covariate, along with *smoking* and *alcohol*. This additional covariate is created by multiplying the *smoking* variable (the number of cigarettes smoked) and the *alcohol* variable (the amount of alcohol consumed). As *smoking-alcohol* is derived mathematically from both *smoking* and *alcohol*, there will be substantial correlations amongst the three variables. However, the correlation between *smoking-alcohol* and either *smoking* or *alcohol* could be considerably reduced if the interaction term *smoking-alcohol* was generated after the values of *smoking* and *alcohol* were centred,[9] ie transformed by subtracting the mean values of each from the original variables. For example, suppose there are five patients in a study, and the number of cigarettes smoked per day by each patient is 5, 10, 15, 20, and 25, respectively. After centring, the values for the variable smoking become -10, -5, 0, 5, and 10, since the mean number of cigarettes smoked is 15.

Apart from problems caused by quadratic terms and product interaction terms, the centring of explanatory variables, in general, does not solve the problem of collinearity or multicollinearity because, mathematically, the correlation coefficient can be interpreted as a product term of two centred variables divided by their variances. Thus, unless the problem is caused by collinearity/multicollinearity between only the intercept and other explanatory variables, both the direction of association between the outcome and collinear covariates and all associated significance testing remain unchanged after centring collinear covariates.

### Principal component analysis and ridge regression

Principal component analysis (PCA) has been proposed as a solution to the numerical problems caused by collinearity and multicollinearity.[7,9,10] The explanatory variables are centred and reorganised into uncorrelated components. Each principal component is a linear combination of all explanatory variables, and the number of principal components is equivalent to the number of explanatory variables. Researchers then usually select the first few principal components that explain most of the variance of the covariates, and use multiple regression analysis to regress the outcome on the selected principal components. The regression coefficients of each original explanatory variable are then derived from the regression coefficients of the selected principal components. The advantage of PCA is that, by selecting only a few principal components (ie not all), the problem of wrong signs amongst regression coefficients (ie the sign of regression coefficient being contradictory to expectation) is usually corrected.

However, one important drawback of PCA is that the principal components selected might well explain the variances of the covariates but poorly explain the variance of the outcome.[10,20] Another commonly recommended method by statistical textbooks, though relatively unknown to most dental researchers, is ridge regression.[21] By adding small values to the explanatory variables, this approach provides biased but more stable estimates of regression coefficients.[10,15,21,22] It should also be noticed that PCA and ridge regression are of no use if there exists perfect collinearity or multicollinearity within one's data.

As these two methods involve advanced statistical theory and complex mathematical computations, detailed descriptions of these methods are beyond the scope of this article, and we strongly recommend that dental researchers consult professional statisticians before embarking upon such complex analyses.

## CONCLUSION

Multivariable regression analyses are useful tools for oral health research, but only if users properly understand their underlying assumptions and limitations. Although multivariable analysis has been used widely, more effort is needed to improve basic understanding of these complex statistical methods amongst oral health researchers. Regression diagnostics for collinearity should be adopted and reported by studies in which complex regression models are used. We strongly suggest that dental researchers consult professional biostatisticians with experience of statistical modelling of clinical data (often collinear), and avoid embarking upon complex statistical analyses themselves.


1. Altman D G. Statistics in medical journals: developments in the 1980s. *Statistics in Medicine* 1991; **10:** 1897-1913.
2. Altman D G. Statistics in medical journals. *Statistics in Medicine* 1982; **1:** 59-71.
3. Tu Y, Gilthorpe M S, Griffiths G S. Is reduction of pocket probing depth correlated with the baseline value or is it 'mathematical coupling'? *J Dent Res* 2002; **81:** 722-726.
4. Tu Y-K, Maddick I H, Griffiths G S, Gilthorpe M S. Mathematical coupling still undermines the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration. *J Dent* 2004; **32:** 133-142.
5. Tu Y-K, Clerehugh V, Gilthorpe M S. Ratio variables in regression analysis can give rise to spurious results: a lesson from guided tissue regeneration. *J Dent* 2004; **32:** 143-151.
6. Miles J, Shelvin M. *Applying regression and correlation*. London: Sage Publication, 2001.
7. Glantz S A, Slinker B Y. *Applied regression and analysis of variance*. New York: McGraw-Hill, 2001.
8. Pedhazur E J. *Multiple regression in behavioral research: Explanation and prediction*. pp 294-313. Fort Worth: Harcourt, 1997.
9. Slinker B Y, Glantz S A. Multiple regression for physiological data analysis: the problem of multicollinearity. *Amer J Phys* 1985; **249:** R1-R12.
10. Chatterjee S, Hadi A S, Price B. *Regression analysis by example*. 3rd Ed. pp 225-284. New York: John Wiley & Sons, 2000.
11. Maddala G S. *Introduction to econometrics*. 3rd Ed. pp 267-300.Chichester: John Wiley & Sons, 2001.
12. Kirkwood B, Stern J A C. *Essential medical statistics*. 2nd Ed. pp 337-339. Oxford: Blackwell, 2003.
13. Moles D. Further statistics in dentistry: Introduction. *Br Dent J* 2002; **193:** 375.
14. Petrie A, Bulman J S, Osborn J F. Further statistics in dentistry Part 6: Multiple linear regression *Br Dent J* 2002; **193:** 675-682.
15. Tu Y-K, Clerehugh V, Gilthorpe MS. Collinearity in linear regression is a serious problem in oral health research. *Euro J Oral Sci* 2004; **112:** 389-397.
16. Draper N R, Smith H. *Applied regression analysis*. 3rd Ed. New York: John Wiley & Sons, 1998.
17. Seki M, Karakama F, Terajima T, Ichikawa, Ozaki T, Yoshida S, Yamashita Y. Evaluation of *Mutans Streptococci* in plaque and saliva: correlation with caries development in preschool children. *J Dent* 2003; **31:** 283-290.
18. Jansson L, Lavstedt S, Zimmerman M. Prediction of marginal bone loss and tooth loss — a prospective study over 20 years. *J Clin Perio* 2002; **29:** 672-678.
19. Simonpietri J J, Novaes A B, Batista E L Jr, Feres Filho E J. Guided tissue regeneration associated with bone-derived anorganic bone in mandibular class II furcation defects. 6-month results at re-entry. *J Perio* 2000; **71:** 904-911.
20. Hadi A S, Ling R F. Some cautionary notes on the use of principle components regression. *American Statistician* 1998; **52:** 15-19.
21. Hoerl A E, Kennard R W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12:** 69-82.
22. Belsley D A. *Conditioning diagnostics*. New York: John Wiley & Sons, 1991.