

## On the inappropriateness of stepwise regression analysis for model building and testing

Moh H. Malek · Dale E. Berger · Jared W. Coburn

Accepted: 30 April 2007 / Published online: 23 May 2007  
© Springer-Verlag 2007

Dear Editor-in-Chief Pietro E. di Prampero,

My colleagues and I recently read the study by Sanada et al. (2007) on the development of prediction models for maximal oxygen uptake ( $\dot{V}O_{2\max}$ ). We would like to focus our comments on two particular areas of the study: (1) the use of stepwise regression, and (2) the practical application of the prediction equations.

1. Use of stepwise regression: Stepwise regression allows a computer program to select a small set of the ‘best’ predictors from a larger set of potential predictors (Tabachnick and Fidell 2001). Stepwise procedures should not be used to develop prediction models because this method produces an inflated  $R$ -squared ( $R^2$ ), inaccurate test of statistical significance, and it does not maximize the theoretical or practical value of the model (Berger 2004; Keppel and Wickens 2004). An essential problem is that estimates of population multiple correlations and tests of statistical significance fail to take into account how many variables were considered in the stepwise analysis. Inflation occurs whether the experimenter selects predictors after looking at the correlations or stepwise regression is used to select the ‘best’ predictors out of a larger set of

potential predictors (Cohen et al. 2003). A more realistic estimate of the population multiple correlation is ‘shrunk’  $R^2$  based on the total number of variables considered. In the Sanada et al. (2007) study where the two strongest predictors from a set of 15 potential predictors produced  $R^2$  of 0.72 with a sample of  $N = 40$ , an estimate of the population multiple  $R^2$  based on 15 predictors is the shrunken  $R^2$  of 0.55. Contrary to the conclusions of Sanada et al. (2007) based on their inflated  $R^2$ , their model offers no improvement on models generated in larger studies as shown in their Table 5.

Ordinarily a regression formula generated on one sample will produce a smaller  $R^2$  when it is applied to a new sample (Pedhazur 1997). Thus, it is surprising that Sanada et al. (2007) found  $R^2$  to be larger ( $R^2 = 0.83$ ) in the validation group than in the derivation group ( $R^2 = 0.72$ ) for which the model was generated. Perhaps this can be explained by large sampling error due to the extremely small sample size ( $N = 20$ ) for the validation group.

In practice, it is always preferable for the investigator to control the order of entry of predictor variables based on theoretical considerations (Berger 2004). This procedure is called ‘‘hierarchical analysis,’’ and it requires the investigator to plan the analysis with care, prior to looking at the data. The double advantage of hierarchical methods over stepwise methods is that there is less capitalization on chance, and careful choice of the order of entry of predictors assures that results such as  $R^2$  added are maximally interpretable (Berger 2004). Kerlinger (1986) stated that, ‘‘... the research problem and the theory behind the problem should determine the order of entry of variables in multiple regression analysis.’’ (p. 545). For example, Malek et al. (2004b, 2005) used hierarchical analysis to develop nonexercise-based

---

M. H. Malek (✉)

Human Performance Laboratory, Department of Nutrition and Health Sciences, University of Nebraska-Lincoln,  
110 Ruth Leverton Hall, Lincoln, NE 68583-0806, USA  
e-mail: mmalek@unlserve.unl.edu

D. E. Berger

Department of Psychology, Claremont Graduate University,  
Claremont, CA 91711, USA

J. W. Coburn

Division of Kinesiology, California State University, Fullerton,  
Fullerton, CA 92834-6870, USA

$\dot{V}O_{2\max}$  prediction model for aerobically trained men and women. The investigators purposefully controlled the order of entry for their predictor variables in order to determine the contribution of physical activity indices (i.e., duration, intensity of the exercise, and the length of time subjects performed habitual physical activity) on  $\dot{V}O_{2\max}$  above and beyond traditional predictors such as age, height, and weight. The comment by Sanada et al. (2007) that, “Although they (Malek et al. 2004b) suggested that the prediction equation is a valid method, their study used six predictors and no statistical selection, such as stepwise regression analysis ...” (p.147) reflects a common misunderstanding of the pitfalls of using stepwise regression. Stepwise regression is entirely data driven, tests of statistical significance reported by popular statistics programs are incorrect, and the set of predictors identified by stepwise regression may not be ‘best’ in terms of generalizability or in terms of theoretical or practical value.

2. Practicality of the Sanada et al. equations: Although we recognize and appreciate the efforts of Sanada et al. (2007) in conducting their study, the predictor variables (thigh skeletal muscle and stroke volume) that are used in their model are not readily available to the general population. Many more practical  $\dot{V}O_{2\max}$  prediction equations derived with larger samples are already available. For example, formulas exist in the Exercise Science literature for treadmill and cycle ergometry (Vehrs et al. 2007; Malek et al. 2004a) as well as for various populations including adult men and women (Storer et al. 1990), teenage athletes (Wells et al. 1973), college students (George et al. 1997), older adults (Blackie et al. 1989), healthy Malaysian and Indian men (Singh et al. 1989; Verma et al. 1998), and aerobically trained individuals (Malek et al. 2004b, 2005).

Sanada et al. (2007) concluded that their results “... suggest that the thigh SM mass and cardiac dimensions are important determinants of  $\dot{V}O_{2\max}$  in healthy young men.” (p.147). This is not a new finding. Wagner’s papers on the determinants of  $\dot{V}O_{2\max}$  and the integrative approach to  $\dot{V}O_{2\max}$  elegantly address the question of which physiological variables contribute to  $\dot{V}O_{2\max}$  (Wagner 1988, 1993, 1996). Therefore, our conclusion is that the Sanada et al. (2007) paper does not add novel information to the  $\dot{V}O_{2\max}$  literature.

## References

- Berger DE (2004) Using regression analysis. In: Wholey JS, Hatry HP, Newcomer KE (eds) Handbook of practical program evaluation. Wiley, San Francisco, pp 479–505
- Blackie SP, Fairbairn MS, McElvaney GN, Morrison NJ, Wilcox PG, Pardy RL (1989) Prediction of maximal oxygen uptake and power during cycle ergometry in subjects older than 55 years of age. *Am Rev Respir Dis* 139:1424–1429
- Cohen J, Cohen P, West S, Aiken L (2003) Applied multiple regression/correlation analysis for the behavioral sciences. L. Erlbaum Associates, Mahwah
- George JD, Stone WJ, Burkett LN (1997) Non-exercise  $VO_{2\max}$  estimation for physically active college students. *Med Sci Sports Exerc* 29:415–423
- Keppel G, Wickens TD (2004) Design and analysis: a researcher’s handbook. Pearson Prentice-Hall, Upper Saddle River
- Kerlinger FN (1986) Foundations of behavioral research. Holt, Rinehart and Winston, New York
- Malek MH, Berger DE, Housh TJ, Coburn JW, Beck TW (2004a) Validity of  $VO_{2\max}$  equations for aerobically trained males and females. *Med Sci Sports Exerc* 36:1427–1432
- Malek MH, Housh TJ, Berger DE, Coburn JW, Beck TW (2004b) A new non-exercise based  $VO_{2\max}$  prediction equation for aerobically trained females. *Med Sci Sports Exerc* 36:1804–1810
- Malek MH, Housh TJ, Berger DE, Coburn JW, Beck TW (2005) A new non-exercise-based  $VO_{2\max}$  prediction equation for aerobically trained men. *J Strength Cond Res* 19:559–565
- Pedhazur EL (1997) Multiple regression in behavioral research. Prentice-Hall, Upper Saddle River, pp 207–211
- Sanada K, Midorikawa T, Yasuda T, Kearns CF, Abe T (2007) Development of nonexercise prediction models of maximal oxygen uptake in healthy Japanese young men. *Eur J Appl Physiol* 99:143–148
- Singh R, Singh HJ, Sirisinghe RG (1989) Cardiopulmonary fitness in a sample of Malaysian population. *Jpn J Physiol* 39:475–485
- Storer TW, Davis JA, Caiozzo VJ (1990) Accurate prediction of  $VO_{2\max}$  in cycle ergometry. *Med Sci Sports Exerc* 22:704–712
- Tabachnick BG, Fidell LS (2001) Using multivariate statistics. Allyn and Bacon, Boston
- Vehrs PR, George JD, Fellingham GW, Plowman SA, Dustman-Allen K (2007) Submaximal treadmill exercise test to predict  $VO_{2\max}$  in fit adults. *Meas in Phys Educ Sci* 11:61–72
- Verma SS, Sharma YK, Kishore N (1998) Prediction of maximal aerobic power in healthy Indian males 21–58 years of age. *Z Morphol Anthropol* 82:103–110
- Wagner PD (1988) An integrated view of the determinants of maximum oxygen uptake. *Adv Exp Med Biol* 227:245–256
- Wagner PD (1993) Algebraic analysis of the determinants of  $VO_{2\max}$ . *Respir Physiol* 93:221–237
- Wagner PD (1996) Determinants of maximal oxygen transport and utilization. *Annu Rev Physiol* 58:21–50
- Wells CL, Scrutton EW, Archibald LD, Cooke WP, De la Mothe JW (1973) Physical working capacity and maximal oxygen uptake of teenaged athletes. *Med Sci Sports* 5:232–241