Research Corner
Regression Analysis for Prediction: Understanding the Process

Phillip B. Palmer, PT, PhD;[1] Dennis G. O'Connell, PT, PhD, CSCS, FASCM[2]

[1]Associate Professor, Hardin-Simmons University, Department of Physical Therapy, Abilene, TX
[2]Professor & Shelton-Lacewell Endowed Chair, Hardin-Simmons University, Department of Physical Therapy, Abilene, TX

ABSTRACT
Research related to cardiorespiratory fitness often uses regression analysis in order to predict cardiorespiratory status or future outcomes. Reading these studies can be tedious and difficult unless the reader has a thorough understanding of the processes used in the analysis. This feature seeks to "simplify" the process of regression analysis for prediction in order to help readers understand this type of study more easily. Examples of the use of this statistical technique are provided in order to facilitate better understanding.

INTRODUCTION

Graded, maximal exercise tests that directly measure maximum oxygen consumption (VO_2max) are impractical in most physical therapy clinics because they require expensive equipment and personnel trained to administer the tests. Performing these tests in the clinic may also require medical supervision; as a result researchers have sought to develop exercise and non-exercise models that would allow clinicians to predict VO_2max without having to perform direct measurement of oxygen uptake. In most cases, the investigators utilize regression analysis to develop their prediction models.

Regression analysis is a statistical technique for determining the relationship between a single dependent (criterion) variable and one or more independent (predictor) variables. The analysis yields a predicted value for the criterion resulting from a linear combination of the predictors. According to Pedhazur,[15] regression analysis has 2 uses in scientific literature: prediction , including classification, and explanation. The following provides a brief review of the use of regression analysis for prediction. Specific emphasis is given to the selection of the predictor variables (assessing model efficiency and accuracy) and cross-validation (assessing model stability). The discussion is not intended to be exhaustive. For a more thorough explanation of regression analysis, the reader is encouraged to consult one of many books written about this statistical technique (eg, Fox;[5] Kleinbaum, Kupper, & Muller;[12] Pedhazur;[15] and Weisberg[16]). Examples of the use of regression analysis for prediction are drawn from a study by Bradshaw et al.[3] In this study, the researchers' stated purpose was to develop an equation for prediction of cardiorespiratory fitness (CRF) based on non-exercise (N-EX) data.

Address correspondence to: Phillip B. Palmer, Hardin-Simmons University, Department of Physical Therapy, Box 16065, Abilene, TX 79698-6065 (ppalmer@hsutx.edu).

SELECTING THE CRITERION (OUTCOME MEASURE)

The first step in regression analysis is to determine the criterion variable. Pedhazur[15] suggests that the criterion have acceptable measurement qualities (ie, reliability and validity). Bradshaw et al[3] used VO_2max as the criterion of choice for their model and measured it using a maximum graded exercise test (GXT) developed by George.[6] George [6] indicated that his protocol for testing compared favorably with the Bruce protocol in terms of predictive ability and had good test-retest reliability (*ICC* = .98 - .99). The American College of Sports Medicine indicates that measurement of VO_2max is the "gold standard" for measuring cardiorespiratory fitness.[1] These facts support that the criterion selected by Bradshaw et al[3] was appropriate and meets the requirements for acceptable reliability and validity.

SELECTING THE PREDICTORS: MODEL EFFICIENCY

Once the criterion has been selected, predictor variables should be identified (model selection). The aim of model selection is to minimize the number of predictors which account for the maximum variance in the criterion.[15] In other words, the most efficient model maximizes the value of the coefficient of determination (R^2). This coefficient estimates the amount of variance in the criterion score accounted for by a linear combination of the predictor variables. The higher the value is for R^2, the less error or unexplained variance and, therefore, the better prediction. R^2 is dependent on the multiple correlation coefficient (R), which describes the relationship between the observed and predicted criterion scores. If there is no difference between the predicted and observed scores, R equals 1.00. This represents a perfect prediction with no error and no unexplained variance (R^2 = 1.00). When R equals 0.00, there is no relationship between the predictor(s) and the criterion and no variance in scores has been explained (R^2 = 0.00). The chosen variables cannot predict the criterion. The goal of model selection is, as stated previously, to develop a model that results in the highest estimated value for R^2.

According to Pedhazur,[15] the value of *R* is often overestimated. The reasons for this are beyond the scope of this discussion; however, the degree of overestimation is affected by sample size. The larger the ratio is between the number of predictors and subjects, the larger the overestimation. To account for this, sample sizes should be large and there should be 15 to 30 subjects per predictor.[11,15] Of course, the most effective way to determine optimal sample size is through statistical power analysis.[11,15]

Another method of determining the best model for prediction is to test the significance of adding one or more variables to the model using the *partial F-test*. This process, which is further discussed by Kleinbaum, Kupper, and Muller,[12] allows for exclusion of predictors that do not contribute significantly to the prediction, allowing determination of the most efficient model of prediction. In general, the *partial F-test* is similar to the *F-test* used in analysis of variance. It assesses the statistical significance of the difference between values for R^2 derived from 2 or more prediction models using a subset of the variables from the original equation. For example, Bradshaw et al[3] indicated that all variables contributed significantly to their prediction. Though the researchers do not detail the procedure used, it is highly likely that different models were tested, excluding one or more variables, and the resulting values for R^2 assessed for statistical difference.

Although the techniques discussed above are useful in determining the most efficient model for prediction, theory must be considered in choosing the appropriate variables. Previous research should be examined and predictors selected for which a relationship between the criterion and predictors has been established.[12,15]

It is clear that Bradshaw et al[3] relied on theory and previous research to determine the variables to use in their prediction equation. The 5 variables they chose for inclusion--gender, age, body mass index (BMI), perceived functional ability (PFA), and physical activity rating (PA-R)--had been shown in previous studies to contribute to the prediction of VO_2max (eg, Heil et al;[8] George, Stone, & Burkett[7]). These 5 predictors accounted for 87% ($R = .93$, $R^2 = .87$) of the variance in the predicted values for VO_2max. Based on a ratio of 1:20 (predictor:sample size), this estimate of *R* , and thus R^2, is not likely to be overestimated. The researchers used changes in the value of R^2 to determine whether to include or exclude these or other variables. They reported that removal of perceived functional ability (PFA) as a variable resulted in a decrease in *R* from .93 to .89. Without this variable, the remaining 4 predictors would account for only 79% of the variance in VO_2max. The investigators did note that each predictor variable contributed significantly ($p < .05$) to the prediction of VO_2max (see above discussion related to the *partial F-test)*.

ASSESSING ACCURACY OF THE PREDICTION

Assessing accuracy of the model is best accomplished by analyzing the standard error of estimate (*SEE*) and the percentage that the *SEE* represents of the predicted mean (*SEE %*). The *SEE* represents the degree to which the predicted scores vary from the observed scores on the criterion measure, similar to the standard deviation used in other statistical procedures. According to Jackson,[10] lower values of the *SEE* indicate greater accuracy in prediction. Comparison of the *SEE* for different models using the same sample allows for determination of the most accurate model to use for prediction. *SEE %* is calculated by dividing the *SEE* by the mean of the criterion (*SEE*/mean criterion) and can be used to compare different models derived from different samples.

Bradshaw et al[3] report a *SEE* of 3.44 $mL \cdot kg^{-1} \cdot min^{-1}$ (approximately 1 MET) using all 5 variables in the equation (gender, age, BMI, PFA, PA-R). When the PFA variable is removed from the model, leaving only 4 variables for the prediction (gender, age, BMI, PA-R), the *SEE* increases to 4.20 $mL \cdot kg^{-1} \cdot min^{-1}$. The increase in the error term indicates that the model excluding PFA is less accurate in predicting VO_2max. This is confirmed by the decrease in the value for *R* (see discussion above). The researchers compare their model of prediction with that of George, Stone, and Burkett,[7] indicating that their model is as accurate. It is not advisable to compare models based on the *SEE* if the data were collected from different samples as they were in these 2 studies. That type of comparison should be made using *SEE %*. Bradshaw and colleagues[3] report *SEE %* for their model (8.62%), but do not report values from other models in making comparisons.

Some advocate the use of statistics derived from the predicted residual sum of squares (*PRESS*) as a means of selecting predictors.[2,4,16] These statistics are used more often in cross-validation of models and will be discussed in greater detail later.

ASSESSING STABILITY OF THE MODEL FOR PREDICTION

Once the most efficient and accurate model for prediction has been determined, it is prudent that the model be assessed for stability. A model, or equation, is said to be "stable" if it can be applied to different samples from the same population without losing the accuracy of the prediction. This is accomplished through cross-validation of the model. Cross-validation determines how well the prediction model developed using one sample performs in another sample from the same population. Several methods can be employed for cross-validation, including the use of 2 independent samples, split samples, and *PRESS*-related statistics developed from the same sample.

Using 2 independent samples involves random selection of 2 groups from the same population. One group becomes the "training" or "exploratory" group used for establishing the model of prediction.[5] The second group, the "confirmatory" or "validatory" group is used to assess the model for stability. The researcher compares R^2 values from the 2 groups and assessment of "shrinkage," the difference between the two values for R^2, is used as an indicator of model stability. There is no rule of thumb for interpreting the differences, but Kleinbaum, Kupper, and Muller[12] suggest that "shrinkage" values of

less than 0.10 indicate a stable model. While preferable, the use of independent samples is rarely used due to cost considerations.

A similar technique of cross-validation uses split samples. Once the sample has been selected from the population, it is randomly divided into 2 subgroups. One subgroup becomes the "exploratory" group and the other is used as the "validatory" group. Again, values for R^2 are compared and model stability is assessed by calculating "shrinkage."

Holiday, Ballard, and McKeown[9] advocate the use of PRESS-related statistics for cross-validation of regression models as a means of dealing with the problems of data-splitting. The PRESS method is a jackknife analysis that is used to address the issue of estimate bias associated with the use of small sample sizes.[13] In general, a jackknife analysis calculates the desired test statistic multiple times with individual cases omitted from the calculations. In the case of the PRESS method, residuals, or the differences between the actual values of the criterion for each individual and the predicted value using the formula derived with the individual's data removed from the prediction, are calculated. The PRESS statistic is the sum of the squares of the residuals derived from these calculations and is similar to the sum of squares for the error (SS_{error}) used in analysis of variance (ANOVA). Myers[14] discusses the use of the PRESS statistic and describes in detail how it is calculated. The reader is referred to this text and the article by Holiday, Ballard, and McKeown[9] for additional information.

Once determined, the PRESS statistic can be used to calculate a modified form of R^2 and the *SEE*. R^2_{PRESS} is calculated using the following formula: $R^2_{PRESS} = 1 - [PRESS / SS_{total}]$, where SS_{total} equals the sum of squares for the original regression equation.[14] Standard error of the estimate for PRESS (SEE_{PRESS}) is calculated as follows: $SEE_{PRESS} =$, where *n* equals the number of individual cases.[14] The smaller the difference between the 2 values for R^2 and *SEE*, the more stable the model for prediction. Bradshaw et al[3] used this technique in their investigation. They reported a value for R^2_{PRESS} of .83, a decrease of .04 from R^2 for their prediction model. Using the standard set by Kleinbaum, Kupper, and Muller,[12] the model developed by these researchers would appear to have stability, meaning it could be used for prediction in samples from the same population. This is further supported by the small difference between the *SEE* and the SEE_{PRESS}, 3.44 and 3.63 mL·kg^{-1}·min^{-1}, respectively.

COMPARING TWO DIFFERENT PREDICTION MODELS

A comparison of 2 different models for prediction may help to clarify the use of regression analysis in prediction. Table 1 presents data from 2 studies and will be used in the following discussion.

As noted above, the first step is to select an appropriate criterion, or outcome measure. Bradshaw et al[3] selected VO_2max as their criterion for measuring cardiorespiratory fitness. Heil et al[8] used VO_2peak. These 2 measures are often considered to be the same, however, VO_2peak assumes that conditions for measuring maximum oxygen consumption were not met.[17] It would be optimal to compare models based on the same criterion, but that is not essential, especially since both criteria measure cardiorespiratory fitness in much the same way.

Table 1. Comparison of Two Non-exercise Models for Predicting CRF

Variables	Heil et al [8] *N* = 374	Bradshaw et al [3] *N* = 100
	β	
Intercept	36.580	48.073
Gender (male = 1, female = 0)	3.706	6.178
Age (years)	0.558	-0.246
Age^2	-7.81E-3	
Percent body fat	-0.541	
Body mass index (kg·m^{-2})		-0.619
Activity code (0 – 7)	1.347	
Physical activity rating (0 –10)		0.671
Perceived functional ability		0.712
	***R* (R^2)**	
	.88 (.77)	.93 (.87)
	SEE	
	4.90 mL·kg^{-1}·min^{-1}	3.44 mL·kg^{-1}·min^{-1}
	SEE %	
	12.7%	8.6%

The second step involves selection of variables for prediction. As can be seen in Table 1, both groups of investigators selected 5 variables to use in their model. The 5 variables selected by Bradshaw et al[3] provide a better prediction based on the values for R^2 (.87 and .77), indicating that their model accounts for more variance (87% versus 77%) in the prediction than the model of Heil et al.[8] It should also be noted that the *SEE* calculated in the Bradshaw[3] model (3.44 mL·kg^{-1}·min^{-1}) is less than that reported by Heil et al[8] (4.90 mL·kg^{-1}·min^{-1}). Remember, however, that comparison of the *SEE* should only be made when both models are developed using samples from the same population. Comparing predictions developed from different populations can be accomplished using the *SEE%*. Review of values for the *SEE%* in Table 1 would seem to indicate that the model developed by Bradshaw et al[3] is more accurate because the percentage of the mean value for VO_2max represented by error is less than that reported by Heil et al.[8] In summary, the Bradshaw[3] model would appear to be more efficient, accounting for more variance in the prediction using the same number of variables. It would also appear to be more accurate based on comparison of the *SEE%*.

The 2 models cannot be compared based on stability of the models. Each set of researchers used different methods for cross-validation. Both models, however, appear to be relatively stable based on the data presented. A clinician can assume that either model would perform fairly well when applied to samples from the same populations as those used by the investigators.

SUMMARY

The purpose of this brief review has been to demystify regression analysis for prediction by explaining it in simple

terms and to demonstrate its use. When reviewing research articles in which regression analysis has been used for prediction, physical therapists should ensure that the: (1) criterion chosen for the study is appropriate and meets the standards for reliability and validity, (2) processes used by the investigators to assess both model efficiency and accuracy are appropriate, 3) predictors selected for use in the model are reasonable based on theory or previous research, and 4) investigators assessed model stability through a process of cross-validation, providing the opportunity for others to utilize the prediction model in different samples drawn from the same population.

REFERENCES

1. *ACSM's Guidelines for Exercise Testing and Prescription*. 7th ed. Philadelphia, PA: Lippincott Williams and Wilkins; 2006.
2. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York, NY: John Wiley & Sons; 1980.
3. Bradshaw DI, George JD, Hyde A, et al. An accurate VO2max nonexercise regression model for 18-65 year-old adults. *Res Q Exerc Sport*. 2005; 76:426-432.
4. Cook RD, Weisberg S. *Residuals and Influence in Regression*. New York, NY: Chapman and Hall; 1982.
5. Fox J. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: SAGE Publications; 1997.
6. George JD. Alternative approach to maximal exercise testing and VO_2max prediction in college students. *Res Q Exerc Sport*. 1996;67:452-457.
7. George JD, Stone WJ, Burkett LN. Non-exercise VO_2max estimation for physically active college students. *Med Sci Sports Exerc*. 1997;415-423.
8. Heil DP, Freedson PS, Ahlquist LE, Price J, Rippe JM. Nonexercise regression models to estimate peak oxygen consumption. *Med Sci Sports Exerc*. 1995:599-606.
9. Holiday DB, Ballard JE, McKeown BC. PRESS-related statistics: regression tools for cross-validation and case diagnostics. *Med Sci Sports Exerc*. 1995:612-620.
10. Jackson AS. Application of Regression Analysis to Exercise Science. In: Safrit MJ, Wood TM, eds. *Measurement Concepts in Physical Education and Exercise Science*. Champaign, IL: Human Kinetics Books; 1989.
11. Kerlinger FN, Pedhazur EJ. *Multiple Regression in Behavioral Research*. New York, NY: Holt, Rinehart and Winston, Inc.; 1973.
12. Kleinbaum DG, Kupper LL, Muller KE. *Applied Regression and Other Multivariable Methods*. 2nd ed. Boston, MA: PWS-KENT Publishing Company; 1988.
13. Mosteller F, Tukey JW. Data Analysis, Including Statistics. In: Lindzey G, Aronson E, eds. *The Handbook of Social Psychology*. Reading, MA: Addison-Wesley Publishing Company; 1968.
14. Myers RH. *Classical and Modern Regression with Applications*. 2nd ed. Pacific Grove, CA: Duxbury Thomson Learning; 1990.
15. Pedhazur EJ. *Multiple Regression in Behavioral Research*. 3rd ed. Fort Worth, TX: Harcourt Brace College Publishers; 1997.
16. Weisberg S. *Applied Linear Regression*. 2nd ed. New York, NY: John Wiley & Sons; 1985.
17. Zeballos RJ, Weisman IM. Behind the scenes of cardiopulmonary exercise testing. *Clin Chest Med*. 1994; 15:193-213.