Croatian Journal of Education
Vol: 14 (4/2012), pages: 917-932
Preliminary communication
Paper submitted: 13th March 2012
Paper accepted: 11th May 2012

Partial Least Squares Regression Analysis: Example of Motor Fitness Data

Ivan Šerbetar
Faculty of Teacher Education, University of Zagreb

Abstract

Based on the research example, the article attempts to describe the partial least squares regression (PLS) as a tool used for modelling the explanatory variables for the prediction of the dependents. The research was carried out on the fitness data of 52 children, nine anthropometric variables were used as predictors, while the dependents were composed of five motor fitness tests. A two-component model was obtained where a small fraction of the dependent variation ($R^2Y = .20$) was explained by predictors ($R^2X=.64$). The Q^2 indicator of the predictive capability of the model was rather low (.16). The main advantages of the PLS were demonstrated: the simultaneous handling of multiple independents and dependents.

***Key words**: anthropometric variables; modelling; motor fitness data prediction; partial least squares regression; PLS*

Introduction

Partial least squares regression (PLS or PLSR) is a method based on the multiple regression and the principal component analysis, invented in 1960s by Herman Wold (Abdie, 2010) for use in chemometrics. In 1990s the method (which has recently become known as *projection to latent structures*) has become widely recognized in education (e.g. Campbell and Yates, 2011, Aunio and Niemivirta, 2010, von Suchodoletz, 2009) and social sciences (e.g. Campbell and Ntobedzi, 2007, Jacobs et al., 2011, Yu-Kang et al., 2010).

PLS has several advantages over multiple regression - it can be used on multicolinear data, a large set of predictive variables can be included and several response variables can be modelled simultaneously (Wold et al., 2001).

PLS, primarily based on very comprehensive tutorials, provided by Geladi and Kowalski (1986) and Abdi (2010), could be described in short as a method grounded in the principal components of X (independent) and Y (dependent) matrices which produce factor scores for X and Y. These new variables, X-scores, usually denoted as *t,* are the predictors of Y and at same time they model X. X-scores are the linear combinations of original X variables estimated with the weights coefficients denoted as *w*. X-scores multiplied by loadings *p* are "good summaries" of X (Wold et al., 2001, p 113). In the same way Y-scores, denoted as *u*, multiplied by the weights *c* summarize Y variables. Both X and Y-scores are computed in such a way as to keep the residuals low, while weights are computed with an intention to maximize the covariance between the responses and the associated factor scores. More technically, matrix X is decomposed into the score matrix T, loading the matrix P and the error matrix E. Similarly, Y matrix is decomposed into the score matrix U, loading the matrix Q and into the error matrix F, both taking the following form:

$$\boldsymbol{X} = \boldsymbol{TP}^T + \boldsymbol{E} \qquad (1)$$

$$\boldsymbol{Y} = \boldsymbol{UQ}^T + \boldsymbol{F} \qquad (2)$$

The elementary graphical representation of the PLS model adapted for the present research is shown in Figure 1.

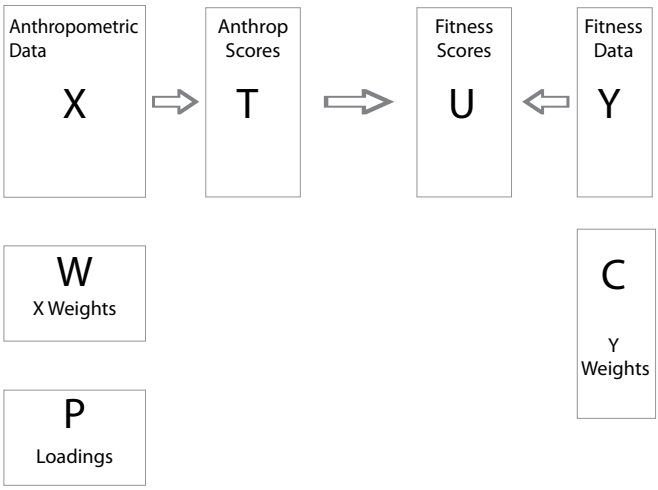

Figure 1. *The PLS model of the present research based on Wold et al. (2001). The factor scores are produced from the initial X and Y data for the final purpose – the prediction of Y from X*

A special issue in PLS analysis is the quality and predictive power of the model. This is usually evaluated by means of resampling procedures like bootstrapping or cross-validation. The latter one implies the *jackknife* technique (*"leave-one-out"*) and V-fold cross-validation. In the jackknife procedure one case is the hold-out when the model is built on others and then tested on that hold-out case; the procedure is repeated n times and each time another case for validation is used.

V-fold cross-validation is a similar method in which the data are divided into a certain number of groups, and parallel models are developed from the reduced data with one of the groups deleted (hold-out; Wold et. al., 2001). When the model is developed, differences between actual and predicted Y values are calculated for the hold-out group. The procedure is repeated for all of the permutations of active and excluded groups and the sum of squares of the above mentioned differences from all the models is used to form the predicted residual sum of squares (PRESS), which is the estimate of the predictive ability of the model (Wold et al., 2001). PRESS is computed as:

$$PRESS = \| Y - \tilde{Y}^{[L]} \| \tag{3}$$

Other coefficients useful for determining the quality of the PLS model are the portion of the explained variation R^2X and the portion of the predicted variation Q^2X. The explained variation is computed as:

$$R^2X = 1 - \frac{residual\ sum\ of\ squares}{sum\ of\ squares} \tag{4}$$

The more significant the principal component is, the higher the R^2X becomes. According to Abdie (2010) the predicted variation is identical to the explained variation except that it is measured on the hold-out fraction of the data and is computed as:

$$Q^2 = 1 - \frac{PRESS_l}{RESS_{l-1}} \tag{5}$$

where RESS is the residual sum of squares of the previous principal component. Q2 statistics also has a practical usage in determining how many components should be retained in the model. The arbitrary threshold value is .0975 (Abdie, 2010).

Methods

Participants and Variables

The data of 52 boys (mean age 8 yrs ± 4 mo.) used in this study have been taken from the author's unpublished children fitness study. The independent variables were composed of anthropometric measures and included: height, weight, abdominal (SK.ABD), subscapular (SK.SUBSC) and triceps (SK.TRIC) skinfold thickness, knee breadth (KNEE.B), thigh circumference (THIGH.C), forearm circumference (FARM.C) and biacromial breadth (BIACRM.B). The motor variables were constituted of the bent arm hang (BentAH), V sit and reach (V-sit), grip strength (GripS), sit-ups (Sit-up), and 3-minute run (Run-3m).

Data Analysis

The data were submitted to the Partial Least Squares analysis performed in STATISTICA 8 employing the NIPALS algorithm. V-fold cross-validation was included and R^2 and Q^2 values were obtained.

The number of components was determined using Q^2 and eigenvalues. During model development, VIP coefficients were computed, and according to the tradition in PLS modelling, the model was presented by means of several graphs. The roll of weight and loadings coefficients in the model was elaborated on as well as the residuals.

Results and Discussion

In the first model developed, only the first component, chosen by cross-validation, appeared to be significant. Since the main purpose of this article was to illustrate the novel research method, which is why a certain depth of the model is expected, the components for the second model were extracted on the basis of eigenvalues greater than 1. Before the second model was built, the contribution to the model of *x* variables was revisited (shown in Figure 2 and Table 1). The importance of the *x* variable for both X and Y in fitting the model is expressed by the VIP coefficient (*variable important for projection; Wold, 1994*). Besides the VIP coefficient with high threshold set by Wold (1994) at 0.8, the predictor weights were checked. The forearm circumference (FARM.C) and the biacromial breadth (BIACRM.B), based on the lowest VIP and weights were removed from the model.

Table 1. *A Partial Least Squares Analysis Summary – 2 components extracted*

C	R^2X	R^2X (Cumul.)	Eigen	R^2Y	R^2Y (Cumul.)	Q^2	Q^2 (Cumul.)	Sign.	Iter.
1	0.37	0.37	3.18	0.16	0.16	0.14	0.14	S	5
2	0.13	0.50	1.03	0.04	0.20	0.03	0.16	S	10

C – components, R^2X (R^2Y) – explained variation in X (Y), Q^2 – predictive quality of the model

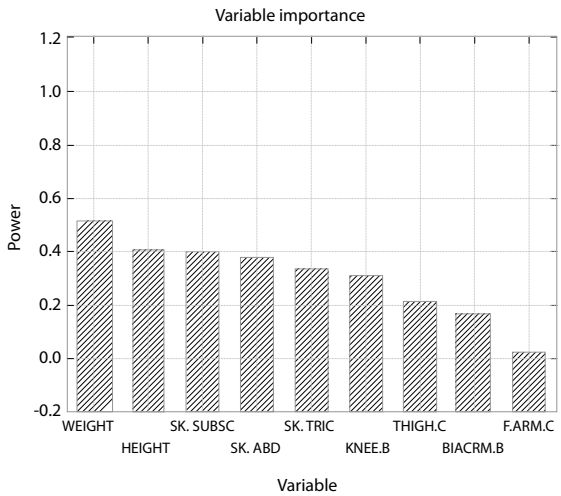

Figure 2. *The important variables for the projection*

The model was rebuilt and a minor decrease in Y variance was observed (Table 2), but a slight improvement of Q^2, with a much larger fraction of the explained

X variation, justified the retention of fewer X variables, especially for reasons of parsimony.

The new model, based on 2 components with eigenvalues larger than 1, accounted for a 65% variation of X which in turn accounted for the 19% variation in Y. The cumulative Q^2 of .17 was obtained.

Table2. *A Partial Least Squares Analysis – the rebuilt model*

C	R^2X	R^2X (Cumul.)	Eigen	R^2Y	R^2Y (Cumul.)	Q^2	Q^2 (Cumul.)	Sign.	Iter.
1	0.47	0.47	3.17	0.16	0.16	0.14	0.14	S	4
2	0.17	0.65	1.12	0.03	0.19	0.03	0.17	S	9

C – components, R^2X (R^2Y) – explained variation in X (Y), Q^2 – predictive quality of the model

For further interpretation of the PLS model, an examination of the plotted scores is required. Factor scores represent the projections of the X and Y data points along the direction of the principal components. The score plot shows the observations in the space of X factors. In order to inspect the irregularities amongst the data, the X-scores were plotted (Figure 3, left-hand side).

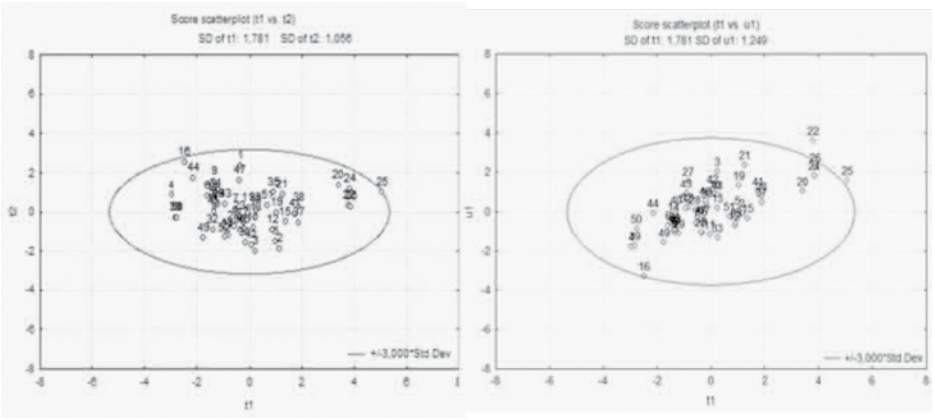

Figure 3. *X scores for the components 1 and 2 (left-hand side), and t scores vs u scores (right-hand side)*

In Figure 3 (left-hand side) X-Scores for the first and second components are plotted. All the observations appear to be consistently clustered into one group without the clearly separated patterns of data. However, there are a few cases with larger positive X-scores spread out on the component 1, appear to be bordering with the outliers. The layout of the X and Y-scores is shown in Figure 3 (right-hand side). The X scores for the first component are associated with increasing values for Y scores, few cases also appear to be outliers. Similarities and dissimilarities of the objects presented in the model can be evaluated through the relations of *t* and *u* scores but also through the orientations of the variables' axes, which is presented in Figure 4.

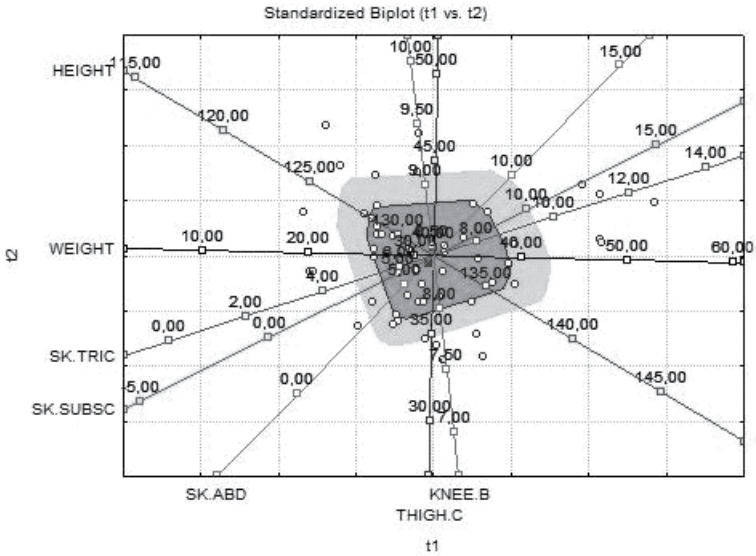

Figure 4. *A standardized biplot showing correlations and orientations of variable axes*

Loadings and weights indicate how much some independent variable contributes to the latent component. X-weights represent the correlation of the independent variables (X) with the scores of the dependent variables (Y), while (X) loadings represent the angle cosines of the direction of the line of best fit.

The combined information of loadings and scores is shown in Figure 4 and reported in Table 3 (loadings). Figure 4 shows how predictors form the space of the latent variables and how they are combined with the observations. Weight is almost entirely lined up with the first component, which means that weight loads heavily on that factor. The positions and orientations of other variables' axes, along with the loading coefficients (Table 3), point out that triceps and subscapular skinfolds are the second and third most important variables which define the first component. Height has less influence on the first component and, like the abdominal skinfold, loads on the second component as well. The proximity of the axes indicates that triceps and subscapular skinfold are highly correlated as well as knee breadth and thigh circumference. The latter two also define the second principal component. Figure 4 also shows how the skinfolds and body weight are close to each other, which means that they also are strongly correlated. On the other hand, height and knee breadth are on the opposite sites, which means that they are negatively correlated and in that way differently affect the model.

Table 3. *X weights and loadings*

Var	XWeight		XLoading	
	Comp1	**Comp2**	**Comp1**	**Comp2**
SK.ABD	0.35	0.47	0.43	0.37
SK.SUBSC	0.44	-0.05	0.49	0.20
SK.TRIC	0.35	0.22	0.45	0.12
WEIGHT	0.57	0.02	0.53	-0.01
HEIGHT	0.44	-0.15	0.36	-0.18
THIGH.C	-0.05	0.51	0.02	0.63
KNEE.B	-0.20	0.67	-0.10	0.70

Like loadings, weights coefficients (Table 4) indicate that body weight, followed by height and subscapular triceps are the most important variables for the definition of the first component, while knee breadth and thigh circumference only influence the second component (see also Figure 5).

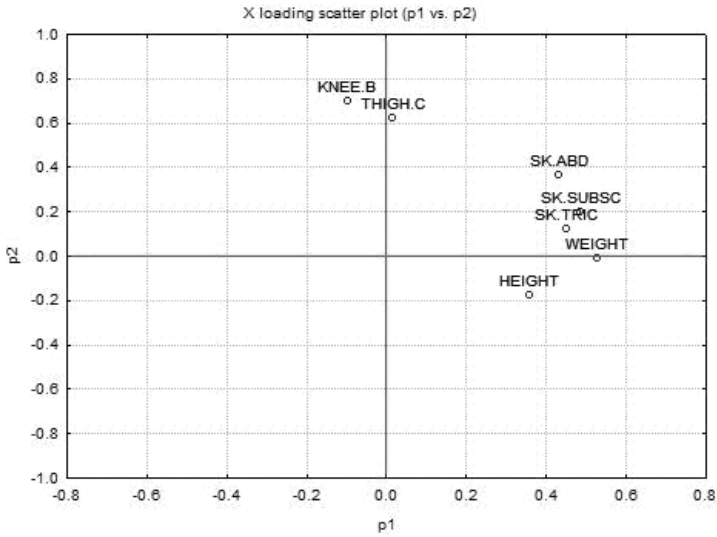

Figure 5. *Loadings of x variables*

According to Hulland (1999) loadings of independent variables should be .7 or higher to confirm that a variable is represented with a specific factor. Raubenheimer (2004) disagrees, stating that the value of .7 is high and real-life data may not reach that standard so, for exploratory purposes, .4 can be used for the central factor and .25 for other factors.

Loadings are supposed to impute meanings for the factors (Table 4), which is rather difficult in the present research because there is not a simple factor structure, namely, cross-loadings of Bent-arm hang and 3 minute running are observed. Both variables are negatively related to the first component, which is expected and related to t-scores. V-sit and reach, and grip strength also load on the first component while the second

component is defined by the positive loadings of bent arm hang, 3-minute running and sit-ups.

Table 4. *Y loadings*

	Comp 1	Comp 2
BentAH	-0.54	0.58
V-sit	0.46	0.01
GripS	0.57	-0.07
Sit-up	-0.10	0.30
Run-3m	-0.40	0.75

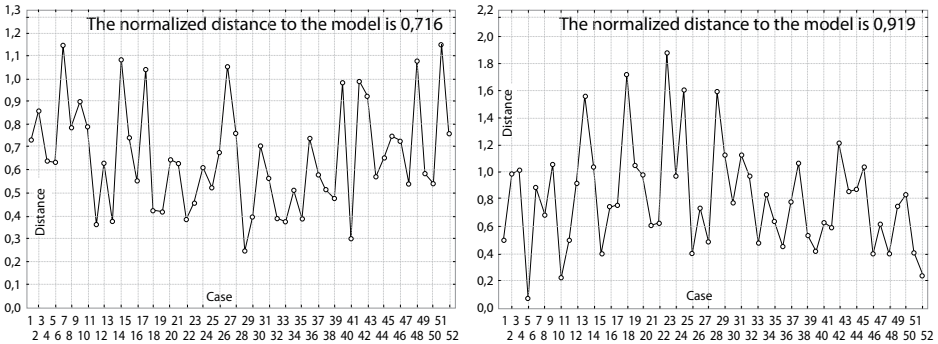

Figure 6. *Distance to the model X (left) and distance to the model Y (right)*

The usual model validation procedure in the PLS requires thorough evaluation of the residuals calculated as the difference between the X data and the prediction of the model, representing the unexplained variation. In other words, residuals represent the fraction of the data not included in the principal components.

At the observation level the basic diagnostics implying the inspection of strong outliers by the t score values has already been demonstrated. Besides that, severe outliers can be detected by Hotelling T^2.

Moderate outliers have a too large Euclidean distance to the model (residual vector module). They are detected via the *Distance to the model* parameter (abbreviated *Dmod* or *D-to-model*) which is the ratio of the absolute and normalized distance to the model. The squared ratio of the absolute and normalized distance to the model approximates Fisher-Snedecor distribution and if the observation is larger than $F_{critical}$ then the observation is a potential outlier (Eriksson et al., 1999). Based on Figure 6, moderate outliers were not detected in the present data.

Conclusion

The intention of this article was to give an empirical overview of an alternative method for the prediction of variables from one set to another. There is similarity between the PLS regression and MLR regression but although both methods deal with linear models, the ways of obtaining regression coefficients are different. The PLS

regression is a method which takes into account the structure of the explanatory and dependent variables simultaneously. Basically, the PLS iteratively decomposes the X and Y matrices in latent structures which are structured from score vectors containing most of the variance in the original X and Y variables. NIPALS algorithm in iterative procedures produces successive orthogonal factors (thus solving the collinearity issue) which maximizes the covariance between the scores.

In the present research, two components were extracted carrying 65% of variation in the predictors, which explained the small size of Y (19%). Consequently, the Q^2 indicator of the predictive capability of the model (Eriksson et al., 1999) was rather low, pointing out that the model is weak, which is expected in this particular case, because the dependent fitness variables could not be predicted solely on the basis of the anthropometric variables.

In this article, which is primarily meant to serve as an example, weight coefficients indicate that the body weight, height and subscapular triceps are the most influential variables for the model. In the final part of the analysis residuals were assessed.

The research has confirmed the already known advantages of the PLS like handling multiple dependents and independents at the same time, the capacity to handle collinearity, etc. As the main disadvantage, difficult interpretation of loadings should be mentioned, because they are the crossproduct of both factor scores. To sum up, the PLS could be used as a strong and effective predictive technique with the less effective interpretative power.

References:

Abdi, H. (2010). Partial least square regression, projection on latent structure regression, PLS-Regression. Wiley Interdisciplinary Reviews: Computational Statistics, 2, 97-106.

Aunio, P., Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. Learning and Individual Difference, 20, 427-435.

Campbell, A., Ntobedzi, A. (2007). Emotional Intelligence, Coping and Psychological Distress: A Partial Least Squares Approach to Developing a Predictive Model. Electronic Journal of Applied Psychology, 3 (1), 39-54. Retrieved February 2.2.2012. from http://jrre.psu.edu/articles/26-4.pdf.

Campbell, A. M., Yates, G. C. R. (2011). Want to be a country teacher? No, I am too metrocentric. Journal of Research in Rural Education, 26(4), 1-12. Retrieved 25.1.2012. from http://jrre.psu.edu/articles/26-4.pdf.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S. (1999). Introduction To Multi and Megavariate data analysis using projection methods. Umetrics AB. Umea, Sweden.

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

Hulland, J. (1999). Use of Partial Least Squares (PLS) in Strategic Management Research: A review of four recent studies. Strategic Management Journal, 20, 195-204.

Jacobs, N., Hagger, M. S., Streukens, S., De Bourdeaudhuij, I., Claes, N. (2011). Testing an integrated model of the theory of planned behaviour and self-determination theory for different energy balance-related behaviours and intervention intensities. British Journal of Health Psychology, 16(1): 113-134.

Raubenheimer, J. E. (2004). An item selection procedure to maximize scale reliability and validity. South African Journal of Industrial Psychology, 30 (4), 59-64.

von Suchodoletz, A., Trommsdorff, G., Heikamp, T., Wieber, F., Gollwitzer, P. M. (2009). Transition to school: The role of kindergarten children's behavior regulation. Learning and Individual Differences, 19, 561-566.

Wold, H. (1994). PLS for multivariate linear modeling. In H. van der Waterbeemd, ed., QSAR: Chemometric methods in molecular design: Methods and principles in medicinal chemistry. Weinheim, Germany: Verlag-Chemie.

Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58, 109-130.

Yu-Kang, T., Woolston, A., Baxter, P.D., Gilthorpe, M.S. (2010). Assessing the Impact of Body Size in Childhood and Adolescence on Blood Pressure: An Application of Partial Least Squares Regression. Epidemiology, 21-4, 440-448.

Ivan Šerbetar
Faculty of Teacher Education, University of Zagreb
Dr. Ante Starčevića 55, 40 000 Čakovec, Croatia
ivan.serbetar@ufzg.hr

Parcijalna regresija metodom najmanjih kvadrata: primjer izveden na podatcima iz motoričkog fitnesa

Sažetak

U ovom se radu, na temelju primjera istraživanja, nastoji opisati parcijalna regresija metodom najmanjih kvadrata (PLS) kao metoda za modeliranje eksplanatornih varijabli u predviđanju zavisnih varijabli. Analiza je izvedena na podatcima 52 djeteta, uključujući devet antropometrijskih varijabli koje su predstavljale prediktore, dok su zavisne varijable bile sastavljene od pet testova motoričkog fitnesa. Dobiven je model s dvije komponente u kojem su prediktori (R^2X= .64) objasnili mali dio varijabilnosti u zavisnim varijablama (R^2Y = .20), pokazatelj kvalitete predikcije modela Q^2 je bio nizak (.16). U istraživanju je prikazana glavna prednost PLS metode: istodobno uključivanje nekoliko nezavisnih i zavisnih varijabli.

***Ključne riječi**: antropometrijske varijable; modeliranje; parcijalna regresija metodom najmanjih kvadrata; PLS; predikcija motoričkog fitnesa*

Uvod

Parcijalna regresija metodom najmanjih kvadrata (PLS ili PLSR) je metoda zasnovana na višestrukoj regresiji i analizi glavnih komponenata, koju je tijekom šezdesetih godina prošlog stoljeća ustanovio Herman Wold (Abdie, 2010) za uporabu u kemometrici. Metoda, za koju se u zadnje vrijeme koristi i naziv *projekcija na latentne strukture*, posljednjih je godina prepoznata i u edukacijskim disciplinama (npr. Campbell i Yates, 2011, Aunioi Niemivirta, 2010, von Suchodoletz, 2009) te općenito u društvenim znanostima (npr. Campbell i Ntobedzi, 2007, Jacobs i sur., 2011, Yu-Kang i sur., 2010).

PLS ima nekoliko prednosti u odnosu na multiplu regresiju – može se koristiti na multikolinearnim podatcima, može uključivati veliki skup nezavisnih varijabli, a osim toga nekoliko zavisnih varijabli može se modelirati istovremeno (Wold i sur., 2001).

Bazirano na iscrpnim tutorijalima Geladija i Kowalskog (1986) te Abdija (2010), PLS se može kratko opisati kao metoda zasnovana na glavnim komponentama matrica X (nezavisnih) i Y (zavisnih varijabli) koje produciraju faktorske skorove za X i Y.

Nove varijable, X-skorovi, koji se obično označavaju *t*, su prediktori za Y, a istovremeno modeliraju X.X-skorovi su linearne kombinacije originalnih X varijabli procijenjenih pomoću regresijskih pondera označenih s *w*. X-skorovi, pomnoženi s faktorskim opterećenjima *p* predstavljaju „dobar sažetak" X sustava (Wold i sur., 2001, p 113.). Istovremeno, Y-skorovi, označeni s *u*, pomnoženi s *c* ponderima prikazuju Y varijable. Skorovi X i Y se izračunavaju tako da se zadrže niski reziduali, dok se regresijski ponderi izračunavaju na način da maksimiziraju kovarijancu između zavisnih varijabli i faktorskih skorova. Tehnički opisano, matrica X se raščlanjuje na matricu skorova T, matricu faktorskih opterećenja P i matricu pogreške E. Slično se raščlanjuje i Y matrica - na matricu skorova U, matricu faktorskih opterećenja Q i matricu pogreške F. Oba dvije poprimaju formu:

$$\boldsymbol{X} = \boldsymbol{TP}^T + \boldsymbol{E} \quad (1)$$

$$\boldsymbol{Y} = \boldsymbol{UQ}^T + \boldsymbol{F} \quad (2)$$

Osnovna grafička reprezentacija PLS modela prilagođenog ovom istraživanju prikazana je na slici 1.

Slika 1.

Posebno pitanje u PLS analizi je kvaliteta i prediktivna snaga modela, što se uobičajeno procjenjuje procedurama ponovnog uzorkovanja kao što su *bootstrap* metoda ili križnavalidacija. Zadnja navedena uključuje i *jackknife* tehniku (*izostavljanje jednog člana uzorka*) i *V-fold* križnu validaciju. U *jackknife* tehnici se jedan član uzorka izostavlja dok se istovremeno model gradi na preostalim članovima i testira na izostavljenom članu, procedura se ponavlja *n* puta, a svaki put se u validaciji koristi drugi član.

V-fold je slična metoda u kojoj se podatci dijele na određeni broj skupina, nakon čega se razvijaju paralelni modeli iz reduciranog uzorka dok je jedna skupina isključena (Wold i sur., 2001). Nakon što se model razvije, izračunavaju se razlike između aktualnih i predviđenih Y vrijednosti za izostavljenu skupinu. Procedura se ponavlja za sve permutacije aktivnih i isključenih skupina, a zbroj kvadrata navedenih razlika iz svih modela se koristi kako bi se formirao predviđeni rezidualni zbroj kvadrata (PRESS) koji je procjena prediktivne sposobnosti modela (Wold et al., 2001). PRESS se izračunava iz:

$$PRESS = \| Y - \tilde{Y}^{[L]} \|^2 \quad (3)$$

Drugi koeficijenti korisni za određivanje kvalitete PLS modela su R^2X - dio objašnjene varijacije i dio predviđene varijacije Q^2X.

Objašnjena varijacija se izračunava iz:

$$R^2X = 1 - \frac{rezidualni\ zbroj\ kvadrata}{zbroj\ kvadrata} \quad (4)$$

Koliko glavna komponenta ima veći značaj toliko je veći R^2X.

Prema Abdie (2010) predviđena varijacija je istoznačna objašnjenoj varijaciji osim što je izmjerena na izuzetoj frakciji podataka, a izračunava se iz:

$$Q^2 = 1 - \frac{PRESS_l}{RESS_{l-1}} \quad (5)$$

gdje je RESS rezidualni zbroj kvadrata prijašnje glavne komponente. Q^2 statistik također ima praktičnu uporabu u određenju broja komponenata koje mogu biti zadržane u modelu. Arbitrarna vrijednost praga je .0975 (Abdie, 2010).

Metode

Ispitanici i varijable

U ovom istraživanju korišteni su podatci od 52 dječaka (8 god. ± 4 mj.). Podatci potječu iz neobjavljenog istraživanja o fitnesu djece. Nezavisne varijable su bile sastavljene od 9 antropometrijskih mjera, a uključivale su: težinu, visinu, kožne nabore trbuha (SK.ABD), leđa (SK.SUBSC) i tricepsa (SK.TRIC) te dijametar koljena (KNEE.B), opseg natkoljenice (THIGH.C), opseg podlaktice (FARM.C) i širinu ramena (BIACRM.B). Skup motoričkih (zavisnih) varijabli je uključivao izdržaj u visu (BentAH), pretklon raskoračni (V-sit), dinamometrijsku jakost (GripS), podizanje trupa (Sit-up) te trčanje na 3 minute (Run-3m).

Analiza podataka

Podaci su obrađeni parcijalnom regresijom po metodi najmanjih kvadrata izvedenoj u programu Statistica 8 prema NIPALS algoritmu. Provedena je i *V-fold* križna validacija te su dobiveni R^2 i Q^2 koeficijenti. Broj glavnih komponenata je određen prema Q^2 koeficijentu odnosno prema svojstvenim vrijednostima. Tijekom razvoja modela, izračunati su VIP koeficijenti i sukladno tradiciji u PLS modeliranju, model je prikazan pomoću nekoliko grafičkih prikaza. Naglašena je uloga regresijskih pondera i faktorskih opterećenja kao i reziduala.

Rezultati i rasprava

U prvom modelu, koji je razvijenom na bazi križne validacije, značajna je bila samo prva glavna komponenta. S obzirom na to da je glavna svrha rada ilustracija nove metode, zbog čega je poželjan bogatiji model, u drugom su modelu na temelju svojstvenih vrijednosti većih od jedan, izabrane dvije glavne komponente. Prije nego je postavljen drugi model, razmotren je doprinos x varijabli modelu (slika 2). Važnost *x* varijabli za X i Y prostor u usklađivanju modela se izražava pomoću VIP koeficijenta (*variable important for projection; Wold, 1994*). Osim VIP koeficijenta, koji ima visoko postavljeni prag od 0.8 (Wold, 1994), provjerene su i vrijednosti regresijskih pondera, tako su opseg podlaktice i širina ramena isključeni iz modela što je bazirano na najnižim VIP vrijednostima i regresijskim koeficijentima.

Tablica 1.

Slika 2.

Nakon što je model postavljen ponovno, uočeno je blago smanjenje u varijanci Y (tablica 2), ali i malo poboljšanje Q^2 koeficijenta i više objašnjene varijacije u X, čime je opravdano zadržavanje manje X varijabli, i to posebno zbog razloga parsimonije modela.

Novi model, utemeljen na 2 komponente sa svojstvenim vrijednostima većim od jedan, objasnio je 65% varijacije u X, što nadalje objašnjava 19% varijacije u prostoru Y. Dobiven je kumulativni Q^2 koeficijent u vrijednosti od .17.

Tablica2.

Za daljnju interpretaciju PLS modela korisni su grafički prikazi faktorskih skorova koji reprezentiraju projekcije X i Y točaka po osima glavnih komponenata. Grafički prikazi (slika 2, lijeva strana) pokazuju opservacije u prostoru X faktora, gdje su u svrhu provjere neregularnosti unutar podataka, prvo prikazani X-skorovi.

Slika 3.

Na slici 3 su prikazani X-skorovi za prvu i drugu komponentu. Primjećuje se da sve opservacije formiraju jedan klaster bez jasno izdvojenih obrazaca podataka. Ipak, primjećuje se nekoliko ispitanika s velikim i pozitivnim X skorovima raspršenih po komponenti jedan, koji izgleda predstavljaju granične atipične vrijednosti (*outlieri*). Kombiniranje X i Y skorova prikazano je na slici 2 (desna strana), gdje se zapaža kako su X skorovi za prvu komponentu povezani sa porastom vrijednosti u Y skorovima, također sa nekoliko potencijalnih *outliera*.

Sličnosti i razlike između objekata u modelu mogu se procijeniti kroz relacije *t* i *u* skorova ali i temeljem orijentacije osovina varijabli što se vidi na slici 4.

Slika 4.

Faktorska opterećenja i regresijski ponderi pokazuju koliko neka nezavisna varijabla pridonosi latentnoj komponenti. X-ponderi predstavljaju korelaciju nezavisnih varijabli (X) i faktorskih skorova zavisnih varijabli (Y), dok (X) faktorska opterećenja predstavljaju kosinuse pravca najboljeg slaganja. Kombinirane informacije faktorskih opterećenja i skorova, predstavljene su na slici 4 i tablici 3. Slika 4 zapravo prikazuje kako varijable formiraju latentni prostor i kako su kombinirane tj. u kakvom su međusobnom odnosu. Varijabla težine je skoro potpuno poravnata sa prvom komponentom, što znači da težina značajno opterećuje taj faktor. Pozicija i orijentacija osi (pravaca) drugih varijabli, zajedno sa koeficijentima faktorskih opterećenja (tablica 3), ukazuju da su kožni nabori tricepsa i lopatice druga i treća najvažnija varijabla koje definiraju prvu komponentu.

Težina je manje utjecajna za prvu komponentu i, poput kožnog nabora na trbuhu, opterećuje i drugu komponentu. Blizina osi ukazuje da su nabor na tricepsu i kožni nabor na lopatici visoko povezani isto kao i dijametar koljena i opseg natkoljenice. Zadnje dvije također definiraju drugu glavnu komponentu. Na slici 4 se vidi kako su i

osi kožnih nabora i tjelesna težina blizu jedne drugima što ukazuje na jaku povezanost tih varijabli. S druge strane, tjelesna visina i dijametar koljena su na suprotnim stranama što znači da su negativno korelirani i da diferencirano utječu na model.

Tablica 3.

Slično kao i koeficijenti faktorskih opterećenja, regresijski ponderi (tablica 4) indiciraju da su tjelesna težina, zajedno s visinom i kožnim naborima lopatice i nadlaktice, najvažnije varijable za definiciju prve komponente, dok su širina (dijametar) koljena i opseg podlaktice jedine varijable koje utječu na drugu komponentu (slika 5).

Slika 5.

Prema Hullandu (1999) koeficijenti faktorskih opterećenja nezavisnih varijabli trebali bi biti .7 ili više kako bi potvrdili da varijablu predstavlja određeni faktor. S navedenim se ne slaže Raubenheimer (2004), navodeći kako je vrijednost od .7 previsoka i kako stvarni podatci ne mogu doseći taj standard, tako da se za eksploratorne namjene može koristiti .4 za glavni faktor, a .25 za ostale. Faktorska opterećenja daju značenje faktorima (tablica 4). U ovom je istraživanju problematično označiti faktore budući da struktura faktora nije jednostavna, naime uočava se kako varijable izdržaj u visu i trčanje na tri minute opterećuju oba faktora. Obadvije navedene varijable su u negativnoj relaciji prema prvoj komponenti, što je očekivano i vidljivo i u *t* skorovima. Pretklon raskoračno i dinamometrijska jakost također koreliraju s prvom komponentom, dok je druga komponenta definirana pozitivnom korelacijom s izdržajem u visu zgibom, sa trčanjem na 3 minute i s podizanjem trupa.

Tablica 4.

Slika 6.

Validacija modela, kakva je uobičajena u PLS-u, zahtijeva temeljitu evaluaciju reziduala koji predstavljaju razliku između podataka X i predikcije modela, što reprezentira neobjašnjeni dio varijacije u modelu. Drugim riječima, reziduali su frakcija podataka koja nije obuhvaćena glavnim komponentama.

Na razini opservacije, osnovna dijagnostika podrazumijeva pregled izraženijih atipičnih vrijednosti i to pomoću *t*-skorova, što je već prikazano. No osim na taj način, veće atipične vrijednosti se mogu otkriti i pomoću Hotelling T^2 testa.

Umjereni outlieri imaju prevelike vrijednosti euklidskih distanci prema modelu. Takve atipične vrijednosti se otkrivaju pomoću - *Dmod ili D-to-model* parametara *(udaljenosti od modela)* koji predstavljaju omjer apsolutne i normalizirane udaljenosti od modela. Kvadrirani omjer apsolutne i normalizirane udaljenosti od modela aproksimira Fisher-Snedecorovu distribuciju, a ako je vrijednost određene opservacije veća od kritičnog F *($F_{critical}$)* tada je ta opservacija potencijalna atipična vrijednost (Eriksson i sur., 1999). S obzirom na sliku 5, takvih vrijednosti ipak nema među podatcima iz ovog rada.

Zaključak

Svrha ovog rada bila je dati empirijski uvid u alternativnu metodu predikcije varijabli iz jednog skupa u drugi. Postoji sličnost između PLS i ML regresije, no iako se obje metode zasnivaju na linearnom modelu, način na koji se dobivaju regresijski koeficijenti je različit. PLS regresija je metoda koja u obzir uzima strukturu eksplanatornih i zavisnih varijabli istovremeno. U osnovi, PLS iterativno raščlanjuje X i Y matricu u latentne strukture koje se sastoje od vektora faktorskih skorova i sadrže većinu varijance iz originalnih X i Y varijabli. NIPALS algoritam u iterativnim procedurama stvara sukcesivne ortogonalne faktore (kojima se rješava problem kolinearnosti) koji maksimaliziraju kovarijancu između skorova. U ovom istraživanju su ekstrahirane dvije komponente koje su nosile 65% varijacije u prediktorima, što je objasnilo mali udio u Y (19%).

Posljedično, Q^2 indikator prediktivnog kapaciteta modela (Eriksson i sur., 1999) bio je nizak, ukazujući na slab model, što je u ovom slučaju i očekivano, jer se zavisne varijable fitnesa ne mogu predvidjeti samo sa antropometrijskim varijablama.

U ovom radu, čija je primarna uloga egzemplarne naravi, regresijski ponderi indiciraju da su tjelesna težina, visina te kožni nabor lopatice i tricepsa najutjecajnije varijable u modelu.

U finalnom dijelu rada procijenjeni su reziduali.

U radu su potvrđene već poznate prednosti PLS metode kao istovremeno korištenje više zavisnih i nezavisnih varijabli, otpornost prema kolinearnosti. Kao glavni problem ove metode, treba spomenuti složenost interpretacije faktorskih opterećenja budući da se radi o krosproduktu faktorskih skorova.

Sumirajući, PLS se može koristiti kao jaka i učinkovita prediktivna metoda, s još uvijek nedovoljnom interpretativnom snagom.