OPLS statistical model versus linear regression to assess sonographic predictors of stroke prognosis

Kianoush Fathi Vajargah\(^1\)
Homayoun Sadeghi-Bazargani\(^2,3\)
Robab Mehdizadeh-Esfanjani\(^4\)
Daryoush Savadi-Oskouei\(^4\)
Mehdi Farhoudi\(^4\)

\(^1\)Department of Statistics, Islamic Azad University, Tehran, North Branch, \(^2\)Neuroscience Research Center, Department of Statistics and Epidemiology, Tabriz University of Medical Sciences, Tabriz, Iran; \(^3\)Public Health Department, Karolinska Institute, Stockholm, Sweden; \(^4\)Neuroscience Research Center, Tabriz University of Medical Sciences, Tabriz, Iran

Abstract: The objective of the present study was to assess the comparable applicability of orthogonal projections to latent structures (OPLS) statistical model vs traditional linear regression in order to investigate the role of trans cranial doppler (TCD) sonography in predicting ischemic stroke prognosis. The study was conducted on 116 ischemic stroke patients admitted to a specialty neurology ward. The Unified Neurological Stroke Scale was used once for clinical evaluation on the first week of admission and again six months later. All data was primarily analyzed using simple linear regression and later considered for multivariate analysis using PLS/OPLS models through the SIMCA P+12 statistical software package. The linear regression analysis results used for the identification of TCD predictors of stroke prognosis were confirmed through the OPLS modeling technique. Moreover, in comparison to linear regression, the OPLS model appeared to have higher sensitivity in detecting the predictors of ischemic stroke prognosis and detected several more predictors. Applying the OPLS model made it possible to use both single TCD measures/indicators and arbitrarily dichotomized measures of TCD single vessel involvement as well as the overall TCD result. In conclusion, the authors recommend PLS/OPLS methods as complementary rather than alternative to the available classical regression models such as linear regression.

Keywords: Prognostic study, trans cranial doppler, partial least squares regression, orthogonal projections to latent structures, multicolinearity

Introduction

Multicolinearity, lower statistical power of study, and missing values are considered major threats to traditional regression models in classical statistics when there are a large number of variables and a small sample size. These various outcomes are common occurrences in brain imaging studies. Some methods have been developed to address these issues. For example, partial least squares (PLS) regression analysis is a known statistical method shown to attenuate the above mentioned problems. At the same time, PLS consists of some limitations such as interpretability problems, multi-component results, and biased coefficients in some situations leading to a greater risk of overlooking real correlations.

Orthogonal projections to latent structures (OPLS) is a recent modification to the PLS regression analysis method. The main idea of the OPLS method is to separate the systematic variation in the X variable into two parts: (1) that which is linearly related to Y, and (2) that which is orthogonal to Y. This gives rise to a much better interpretability. Compared to the traditional principal component analysis and factor analysis, both PLS and OPLS methods are known as supervised statistical methods.
Supervised modeling techniques may be suitable alternatives or complementary options to the classical modeling techniques in that they can manage large numbers of variables for smaller sample sizes and simultaneously be less prone to threats from multicollinearity and missing values.\(^1\) The implication of the OPLS statistical method is gaining interest in various research areas, ranging from its origin in chemometrics in 2006 to its most recent applications in areas such as diversified as public health research.\(^2\) Imaging science (in particular, brain imaging) has become a recent field of application for OPLS modeling.\(^7\)–\(^9\) Trans cranial doppler (TCD) is a diagnostic imaging method used in stroke victims and has also been used in stroke risk prediction.\(^13\)

Considering the limitations of classical regression models in managing large numbers of variables, the research oriented use of TCD produces a large number of variables that needs to be modeled through appropriate statistical methods. Limited knowledge is available regarding applicability of newer statistical methodology alternatives in modeling TCD findings as well as modeling the scale based assessment of stroke prognosis.

Thus, the aim of this study was to assess the applicability of the PLS/OPLS statistical models for investigating the role of TCD findings in order to predict ischemic stroke prognosis in comparison to traditional linear regression analysis.

**Methods**

The present study was conducted on 116 ischemic stroke patients admitted to a specialty neurology ward in Tabriz, Iran. TCD was performed on all patients in the first week of admission. In addition, vessel lesions as well as blood velocity in different vessels were determined. On the first week of admission, the Unified Neurological Stroke Scale (UNSS) function scale was used once for clinical evaluation and subsequently used again once more 6 months later. All data were primarily analyzed using simple linear regression analysis and later considered for multivariate analysis using new supervised models through SIMCA P+12 software, a known software package used for supervised modeling techniques (Umetrics AB, SE-90719, Umeå, Sweden). OPLS is implemented in SIMCA P+12 such that the method is available under the standard PCA and PLS modeling framework. PLS and OPLS were considered to be the multivariate analysis methods. Categorical variables (after being converted into dummy variables) were entered into the model as other dichotomous and continuous measures. PLS was run prior to OPLS; however, only the OPLS results are presented in this study.

Four groups of variables were available to be analyzed. The first group included variables pertaining to demographic characteristics, past medical history, and medical examination. The second group of variables was based on laboratory measurements such as cholesterol. The third group of variables came from TCD of all patients and a binary result from computerized tomography scanning or magnetic resonance imaging for some patients. Lastly, the fourth group of variables consisted of scores from the UNSS function scale. UNSS measured after six months was considered to be the response variable that was predicted by other variables. A total of 147 variables were formatted for SIMCA P12 to be analyzed.

The methods of choice for multivariate analysis were considered to be PLS followed by OPLS. Before fitting the PLS/OPLS models, we transformed all variables with a low min/max ratio or high skewness. Log transformation for all respective variables were used and the R-square and Q-square measures were calculated for overall model assessment. A prediction set of 116 and validation set of 34 were both analyzed for validity assessment. Both coefficient variability analysis of variance and Hotelling’s T2 range plot were used to assess whether fitted models were overall statistically significant. The models were also assessed using a Hotelling’s T2 range plot. The normal probability plot of residuals was used to assess the normal distribution of standardized results and a lack of outliers. To make this report easy to read for clinicians, most of the mathematical and technical information have been omitted. However, details in this regard can be found elsewhere.\(^1\)

This study was conducted in accordance with the ethical standards of the responsible committee of ethics in Tabriz University of Medical Sciences, Tabriz, Iran.

**Results**

Descriptive results

A total of 116 observations were analyzed as a prediction set. The mean age amongst the patients was approximately 62.1 years with a standard deviation of 11.9 years. Seventy-three percent of all participants were male. The mean body weight was 72.5 kg with a standard deviation of 10.4 amongst all patients. No pathologic findings were found in brain computerized tomography scans in 36% of the patients. Right hemiplegia was the most common complaint of patients followed by left hemiplegia and dizziness. The mean hemoglobin and hematocrit counts of all patients were 9.8 g/dl and 44.5%, respectively. The mean blood urea nitrogen and serum creatinine levels of the patients were 21.9 and 0.96 mg/dL, respectively.
Linear regression results
Due to linear regression analysis limitations, it was not possible to model all the measurements of blood velocity in TCD. The main limitation in this regard was the available high multicollinearity and large number of variables. Therefore, only the variables from TCDs were modeled and were arbitrarily dichotomized to show if the blood velocity was abnormal in the TCD of single vessels. The results of the multivariate regression analysis showed that the baseline UNSS, right and left sided middle cerebral artery involvement, as well as right sided anterior cerebral artery involvement were statistically significant predictors of the UNSS score at six months.

OPLS regression analysis
Coefficient variability analysis of variance showed that the overall model was statistically significant. Hotelling’s T2 range plot also showed the models to be acceptable. The normal probability plot of residuals, as shown in Figure 1, depicts the normal distribution of standardized results and lack of outliers. Distribution of residuals was shown to have an acceptable low skewness.

The regression model based on orthogonal projections to latent structures was able to identify 26 variables as statistically significant predictors of stroke prognosis. Figure 2 shows all of the 147 variables that were modeled being sorted based on their beta coefficients. The size of the coefficient represents the change in the response when the variable varies from 0 to 1, in coded units (one standard deviation when the data are scaled to unit variance), while the other variables are kept at their averages. These coefficients refer to the centered and scaled X-data or predictors, with Y (6th month UNSS score) just scaled, but not centered. The coefficients basically express how strongly Y is correlated to the systematic part of each of the X-variables. The loadings plot of the model showing the relationship between the various variables is shown in Figure 3. As can be seen in the figure, having an overall normal TCD, versus an abnormal TCD, is quite correlated with the baseline UNSS score.

It was determined that the applied statistical models had good predictive power of approximately 80%. However, the baseline UNSS score had stronger predictability when compared to imaging predictors.

Discussion
The main outcome of interest for assessing the stroke functional prognosis in the present study was UNSS. It has been widely used to assess stroke prognosis in prognostic studies, or in assessing the efficacy of treatments in stroke clinical trials both in ischemic and hemorrhagic stroke. TCD is both a diagnostic tool in stroke prognosis and has also been used in stroke risk prediction.

In the present study, linear regression analysis results identified some TCD predictors of stroke prognosis, which were confirmed through the OPLS modeling method. Moreover, this methodology appeared to have a higher sensitivity in detecting the predictors of stroke prognosis and had detected several more predictors as well. Using this model, it was possible to use the single TCD measures or indicators, arbitrarily dichotomized measures of TCD single
vessel involvement, as well as the overall TCD result. The inclusion of these variables in one single statistical model can be a major source of methodological concern, especially in the case of linear regression. Even when ignoring this problem, another source of concern associated with using linear regression analysis is the decreased statistical power of the study as encountered by increasing the number of parameters in a linear regression model. Another major concern in using baseline UNSS along with other possible predictors to predict the 6th month UNSS score in linear regression is that both scores are from the same type and possibly share common predictors. Thus, if not adequately cared for, this could be a strong source of multicollinearity.

The use of latent variable based models like PLS and OPLS could be a solution to the problem of multicollinearity and also provide higher statistical power in the study. One main finding in the present study was that the results of a modern complex model confirmed the predictors identified by a parsimonious linear regression model. However, it was shown that other than the predictors...
Table 1 Results of OPLS regression analysis in detecting predictors of stroke prognosis based on the UNSS score measured six months after a stroke attack

<table>
<thead>
<tr>
<th></th>
<th>MV Right</th>
<th>MV Left</th>
<th>PSV Right</th>
<th>PSV Left</th>
<th>PI Right</th>
<th>PI Left</th>
<th>RI Right</th>
<th>RI Left</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Part 1:</strong> crude velocity measures of blood velocity in TCD resulting in statistically significant predictors of the sixth month UNSS score</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>ACA</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>MCA</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Siphon</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VB</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICA (extra cranial)</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICA (intra cranial)</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Part 2:** other statistically significant predictors of a sixth month UNSS score

**Notes:** Clinicians decision on involvement of LMCA, LMCAp and RMCA; left sided hemiplegia or dizziness as chief complaint; pathological findings in MRI, PTT, and serum LDL and HDL; having normal TCD; being a smoker, the reason for starting warfarin, and baseline UNSS score. *Statistically significant P < 0.05.

**Abbreviations:** OPLS, orthogonal projections to latent structures; UNSS, Unified Neurological Stroke Scale; TCD, transcranial doppler; ACA, anterior cerebral artery; MCA, middle cerebral artery; PCA, posterior cerebral artery; PI, Pulsatility index; RI, Resistance index; ICA, Internal carotid artery; LMCA, Left Middle Cerebral Artery; RMCA, Right Middle Cerebral Artery; MV, Mean velocity; PSV, Peak systolic velocity; LMCAp, Left Middle Cerebral Artery (proximal); MRI, magnetic resonance imaging; PTT, partial thromboplastin time; LDL, low-density lipoprotein; HDL, high-density lipoprotein.

determined by simple use of linear regression, several other predictors were identified using the OPLS model. This may be due to the higher power of the study in PLS/OPLS methods when compared to linear regression in classic statistics.

Nevertheless, two main disadvantages of such models should always be considered when using them. First is the higher complexity of such models, and secondly is the fact that these methods are rather new and need to be assessed in a substantial number of studies prior to recommending them as a single statistical methodology for modeling TCD findings or when being used in stroke prognostic studies. Therefore, the authors prefer to recommend them as complementary rather than alternative to the available classical regression models such as linear regression. Regardless of the main objective of this study, baseline UNSS score was a better predictor of stroke prognosis in comparison to the TCD predictors. This is suggestive of a lower prognostic predictive value of the TCD findings. Not specific to TCD, other imaging predictors have also been shown to have lower predictive values when compared to baseline clinical assessments; however, this does not undermine the role of combined clinical imaging prediction sets.

**Conclusion**

In comparison to linear regression, the OPLS model appeared to have higher sensitivity in detecting the predictors of ischemic stroke prognosis and detected several more predictors. The study confirmed possible applicability of supervised models to predict some neurological outcomes.

**Disclosure**

The authors report no conflict of interest in this work.

**References**


