Resumen

Las técnicas de regresión deben aplicarse de acuerdo al campo de análisis, el tipo de datos que se quiere analizar, y sobre todo, la problemática de interés. En la extracción de la información por métodos de espectroscopia, utilizando el Infrarrojo Cercano (NIR), la información posee una enorme cantidad de variables que están correlacionadas. Los métodos de estimación como OLS, GLM y otros, no son eficaces en este tipo de contexto, y se hace imperativo contar con otras técnicas de regresiones especializadas, que puedan solucionar al mismo tiempo la multicolinealidad y reducir el número de variables independientes. En el campo de la Quimiometría, dos modelos de regresión utilizados en estos contextos son la regresión por Componentes Principales (PCR) y por Mínimos Cuadrados Parciales (PLS). Cuando el compuesto de análisis presenta una relación lineal, los resultados suelen ser satisfactorios. El propósito es presentar las dos técnicas de regresión, conceptualmente como a partir de su fundamento estadístico-matemático, y confrontar empíricamente la eficacia en los resultados de predicción. Al poseer una cantidad de datos limitados, se realizan simulaciones para medir la sensibilidad y robustez de los resultados. Se pretende predecir el porcentaje de proteína en la semolina de arroz. Los análisis se llevaron a cabo en SAS versión 9.4, mediante la PROC PLS, PROC SQL, y macro programas para la parte de simulación. En la metodología de análisis, se decidió dividir los datos en dos partes: un *training data set* (estimación del mejor modelo) y un *validation data set* (validación de los resultados). En la determinación del número de componentes en cada técnica de regresión, se utilizaron los estadísticos de PRESS, T^2 de Hotelling y el R^2. En la exploración y verificación de los supuestos del modelo, los resultados sólo presentaron problemas a nivel de los valores extremos: la heterogeneidad, normalidad, y los otros gráficos de diagnóstico no presentan una desviación anormal. En la validación de los modelos finales de regresión, los estadísticos del SEP, REP y gráficos de valores observados contra predichos, confirman mejores ajustes para la regresión por PLS. Las simulaciones de Bootstrap y Montecarlo ratifican una mejor estabilidad en los resultados de la regresión por PLS, concluyéndose como mejor técnica para predecir el porcentaje de proteína en la semolina de arroz.

Abstract

Regression techniques should be applied depending on the field of analysis, the data type, and especially the problem of interest. In the extraction of information by spectroscopic methods, using the Near Infrared (NIR), the information has a great number of variables that are highly correlated. The estimation methods such as OLS, GLM, etc. are not effective in this context. It becomes imperative to have other specialized regression techniques that can simultaneously solve the multicollinearity and reduce the large number of independent variables. Two regression models are used in this context: regressions by Principal Components Analysis (PCA) and Partial Least Squares (PLS). The purpose is to present the two regression techniques, conceptually and from its statistical and mathematical foundation. Empirical effectiveness confronts the prediction results. Since the limited number of data, Bootstrap simulation tests are performed to measure the sensitivity of the results. It is intended to predict the percentage of protein in rice bran. Analyses were carried out in SAS 9.4, by the PROC SQL, and macro programs for simulations tests. In the methodology, it was decided to split the data into two parts: a training data set (best model estimation) and a validation data set (validation of results). PRESS, Hotelling 's T and R Square statistics were used to decide the number of components in each regression technique. Model assessment showed problems only with out-liars: heterogeneity, normal, and other diagnostic plots do not exhibited an abnormal deviation. Regression models validation measurements (SEP, REP and graphics of predicted values against observed values) confirm the best settings for PLS regression. Bootstrap and Monte Carlo simulations as well ratify better stability in the results of the PLS regression, concluding as best technique to predict the percentage of protein in rice bran.