ANÁLISIS ESTADÍSTICO COMPARATIVO DE REGRESIÓN LINEAL MÚLTIPLE Y ANÁLISIS DE COMPONENTES PRINCIPALES EN UNA VARIABLE DE RESPUESTA EN PRESENCIA DE MULTICOLINEALIDAD

COMPARATIVE STATISTICAL ANALYSIS OF MULTIPLE LINEAR REGRESSION AND PRINCIPAL COMPONENTS ANALYSIS ON A RESPONSE VARIABLE IN THE PRESENCE OF MULTICOLINEARITY

Andrea Villegas Guerrero
Tecnológico Nacional de México / IT de Celaya, México
m2203046@itcelaya.edu.mx

Armando Javier Ríos Lara
Tecnológico Nacional de México / IT de Celaya, México
armando.rios@itcelaya.edu.mx

José Antonio Vázquez López
Tecnológico Nacional de México / IT de Celaya, México
antonio.vazquez@itcelaya.edu.mx

Moisés Tapia Esquívias
Tecnológico Nacional de México / IT de Celaya, México
moises.tapia@itcelaya.edu.mx

Recepción: 22/noviembre/2023 **Aceptación:** 27/febrero/2024

Resumen

Dentro de los procesos es de gran importancia analizar el efecto que tienen las distintas variables sobre estos. Las técnicas de regresión lineal múltiple (RLM) y análisis de componentes principales (ACP) permiten analizar e identificar las variables independientes que influyen sobre la variable de respuesta. En este artículo se comparan dos modelos utilizando conjuntos de datos simulados con 15 variables independientes, tomando en cuenta la influencia de la multicolinealidad inducida en los resultados. En un modelo se aplica RLM directamente mientras que en el otro se obtienen los componentes principales para aplicar RLM. Se comparan los estadísticos PRESS, R^2, R^2 ajustada y R^2 de la predicción de los modelos. Los hallazgos de este estudio destacan la importancia de considerar la multicolinealidad al seleccionar y aplicar modelos estadísticos en el análisis de datos. Los resultados arrojan que el modelo de RLM es el modelo óptimo para realizar predicciones de mejor manera tanto para conjuntos de datos sin el efecto de la multicolinealidad como para conjuntos de datos que cuenten con ella.

Palabras clave: ACP, análisis de variables, modelos estadísticos, multicolinealidad, RLM.

Abstract

It is important to analyze the effect that the different variables have on the processes. The techniques of multiple linear regression (RLM) and principal component analysis (PCA) allow us to analyze and identify the independent variables that influence the response variable. In this article, two models are compared using simulated data sets, each set has 15 independent variables, and the influence of induced multicollinearity is considered on the results. In one model RML is applied directly to the data sets while in the other the principal components are obtained followed by RLM. The statistics compared for the models are PRESS, R^2*, adjusted* R^2 *and* R^2 *of prediction. The findings of this study highlight that RLM model is the optimal one to make predictions in a better way both for data sets without the effect of multicollinearity and for data sets that have it.*

Keywords*: MLR, multicollinearity, PCA, statistics models, variable analysis.*

1. Introducción

La selección de técnicas apropiadas para optimizar las variables de respuesta representa un papel fundamental tanto en los campos de investigación como en el área industrial. Entre las herramientas más utilizadas está la regresión lineal múltiple (RLM), y con menor frecuencia de uso el análisis de componentes principales (ACP). Estas técnicas se emplean para modelar las relaciones entre variables independientes y dependientes. Aunque los enfoques de las técnicas difieren, ambos proporcionan herramientas para predecir y comprender las variables de respuesta.

Cuando se tiene un gran volumen de datos, específicamente más allá de tres dimensiones, el análisis se puede tornar complicado. En la industria alimenticia es importante abordar problemas relacionados con la composición química, los efectos del procesamiento y trazabilidad, los cuales cuentan con muchas variables para ser analizadas. Un caso similar pasa en la agricultura, para predecir el rendimiento de algún cultivo se tienen muchas variables climatológicas y de suelo lo que hace que la información sea extensa.

Existen diversas técnicas con las cuales podemos reducir dimensiones y realizar predicciones como regresión parcial de mínimos cuadrados, redes neuronales etc. El propósito de este estudio es llevar acabo un análisis comparativo entre MLR y el uso de análisis de componentes principales seguido de regresión lineal múltiple (ACP+MRL) en el contexto de la predicción de variables de respuesta, considerando la presencia de multicolinealidad en datos generados mediante simulación Montercarlo.

La multicolinealidad puede ser especialmente problemática en escenarios prácticos donde los datos se encuentran interrelacionados, como en el caso de modificaciones nutrimentales y químicas al momento de añadir radiación para la conservación de alimentos o para predecir la calidad en algún producto como los vinos [Guillén-Casla et al, 2011]. En tales casos, la multicolinealidad puede conducir a predicciones inexactas o sesgadas, lo que subraya la importancia de abordar este problema de manera efectiva en el modelado estadístico. El ACP es utilizado para la reducción dimensional explicando la mayor variabilidad posible y las interrelaciones de las variables independientes [Gozá-León et al, 2020].

La metodología implica la generación y procesamiento de datos para el análisis. Mediante simulación Monte Carlo se generan los conjuntos de datos con características específicas. Se aplica RLM y ACP para modelar la relación de variables independientes y la de respuesta. Se evalúan los modelos utilizando pruebas estadísticas y se determinan las fortalezas y limitaciones de cada modelo. Este estudio se presenta como un aporte para la comprensión y elección de técnicas de modelado, destacando la importancia de abordar la multicolinealidad en la predicción de variables de respuesta.

Es de gran importancia considerar las particularidades que cada escenario presenta al seleccionar un modelo de predicción, diversos autores han evaluado modelos para determinar cuál de ellos puede realizar una predicción más asertiva. Vervoot y Wolff [2020] usan ACP y RLM en conjunto, lo que permite predecir las concentraciones en espectros con mayor precisión que la técnica de RLM. Para la predicción del rendimiento del arroz, los modelos propuestos por Banakara et al. [2019], RLM obtuvo mejores resultados tanto en R^2 como en porcentaje de predicción. Otro caso de estudio similar en donde el modelo de ACP+RLM tiene mejor porcentaje de R^2 es para la predicción de suministro esporádico en la cadena de suministro [Ahmed et al., 2020], en donde se comparan cuatro modelos siendo superior el ACP+RLM. Por el contrario, RLM resultó superior en el porcentaje de predicción para describir la generación de desperdicio en el área de salud [Minoglou & Komilis, 2018]. La variabilidad en los resultados previos aumenta la importancia de conocer acerca de la información con la que se trabaja. Este artículo contribuye en la selección de modelos predictivos mediante una comparación entre RLM y ACP+RLM en conjuntos de datos que cuentan con y conjuntos de datos sin multicolinealidad.

2. Métodos

Proceso de generación de datos

Para llevar a cabo esta comparación, se generó un modelo verdadero con 15 variables independientes, coeficientes y error obtenidos de manera aleatoria y con distribución normal mediante simulación Montecarlo.

Para el estudio se simularon con el mismo método, 100 conjuntos de datos con distribución normal y sin multicolinealidad, también se generaron 100 conjuntos de datos con distribución normal y multicolinealidad mediante Matlab. Los conjuntos de datos cuentan con 15 variables independientes y una variable de respuesta. La variable de respuesta es generada a partir de las variables independientes, coeficientes y error aleatorio.

La multicolinealidad es inducida en algunas variables independientes del modelo verdadero. Se seleccionaron variables independientes a las cuales se les aplicaron combinaciones lineales para introducir correlaciones entre ellas. Para este caso se asignaron nuevos valores a cuatro columnas de la matriz de datos, utilizando combinaciones lineales de otras columnas, lo que generó correlaciones significativas entre variables. Con este procedimiento se crea un escenario simulado de datos con multicolinealidad.

Los conjuntos se analizaron para cumplir con características para poder ser utilizados para una regresión lineal múltiple. La normalidad y homogeneidad de varianzas en la variable dependiente Y del modelo para los valores de la variable independiente [Montgomery et al., 2007].

Enfoque de modelado

El estudio cuenta con dos enfoques para el análisis comparativo, la regresión lineal múltiple y el análisis de componentes principales seguida de regresión lineal múltiple.

Regresión Linear Múltiple (RLM). Es una técnica clásica de gran uso en diversas áreas para explorar la relación entre dos o más variables independientes y una variable de respuesta. El modelo está dado por la ecuación 1 [Li et al., 2023].

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (1)$$

Donde Y es la variable dependiente, los coeficientes $\beta, \beta_1 \dots \beta$, las variables independientes se denotan como x_1, $x_2 \dots x_k$ y ϵ el error. Se utiliza la herramienta de regresión stepwisse que es utilizado cuando se cuenta con predictores potenciales, pero no con información suficiente para determinar que variables incluir. Se utiliza Y para obtener los valores P y se decide que variables utilizar [Harrell, 2015].

Análisis de Componentes Principales seguida de Regresión Linear Múltiple *(ACP+RLM):* ACP es una técnica de reducción dimensional. Se aplica para obtener las variaciones más significativas de las variables independientes. Seguido de la obtención de los componentes principales (k), estos se utilizan para realizar la regresión y obtener el modelo de la variable de respuesta [Vervoort, S., & Wolff, M., 2020]. Para seleccionar los componentes principales a retener, se utilizó el criterio de acuerdo con Sanguansat, P. [2012] donde numéricamente se elige la cantidad de componentes que expliquen entre el 70 y 90% y gráficamente se toma el número de componentes que se encuentren en el punto de la curvatura más pronunciada contra una línea recta.

Evaluación del desempeño

Hay varias maneras de medir la multicolinealidad, para este caso se utiliza el factor de inflación (VIF). Un valor alto de VIF, generalmente mayor a 10, es indicador de multicolinealidad [Chan et al., 2022]. Para la comparación del desempeño de ambos enfoques se emplean varias métricas de evaluación: R^2, R^2 ajustada, R^2 de la predicción y la suma de cuadrados de predicción (PRESS).

Softwares estadísticos

- Matlab R2020b: Este software es utilizado dentro de las etapas iniciales para generar los conjuntos de datos mediante simulación Montecarlo con características específicas, las cuales incluyen distribución normal, multicolinealidad y aleatorización por medio de codificación.
- Minitab 19: Es utilizado para etapas posteriores del estudio. Es de gran utilidad para la construcción de los modelos de RLM y ACP+RLM, lo cual permite realizar un análisis comparativo de las métricas de los modelos.

3. Resultados

El modelo verdadero generado por Matlab con las características asignadas se muestra en la ecuación 2.

$$\begin{aligned} Y = &-0.6490x_1 + 1.1812x_2 - 0.7585\, x_3 - 1.1096x_4 - 0.8456\, x_5 - 0.5727\, x_6 \\ &-0.5587\, x_7 + 0.1784\, x_8 - 0.1969\, x_9 + 0.5864\, x_{10} - 0.8519\, x_{11} \\ &+0.8003\, x_{12} - 1.5094\, x_{13} + 0.8759\, x_{14} - 0.2428\, x_{15} + \varepsilon \end{aligned} \quad (2)$$

Para asegurar que los datos fueran normales se realizaron pruebas Anderson-Darling donde los valores fueron positivos para este aspecto. Se realizaron pruebas de corridas de aleatoriedad para determinar que los conjuntos de datos fueran aleatorios, los resultados obtenidos confirmaron que estos fueron simulados correctamente.

A partir de los conjuntos de datos simulados, se compararon los modelos. Para los conjuntos de datos que carecían de multicolinealidad se determinó la R^2, donde para RLM fue de 99.25% y para ACP+RLM de 78.59. En la tabla 1 se presentan un resumen de los resultados obtenidos para ambos modelos. En el modelo de RLM, se obtuvo un alto coeficiente de determinación (R^2), lo que indica que el modelo explica casi en su totalidad la variabilidad de la variable de respuesta. Los valores para R^2 ajustada y R^2 de la predicción también son altos para el modelo de RLM en comparación del otro modelo. El valor de PRESS es bajo, lo cual indica una buena precisión en las predicciones. Para el modelo de ACP+RLM en promedio se tomaron 10.43 componentes. Aunque la R^2 es mucho menor (76.71%), aún es considerado un valor significativo. El valor de R^2 de la predicción es de 74.06%. Sin embargo, el valor de PRESS es sumamente mayor que en el modelo de RLM.

Tabla 1 Resumen de resultados para conjunto de datos sin multicolinealidad.

Modelo	R^2 (%)	R^2 ajustada (%)	R^2 de la predicción (%)	PRESS
RLM	99.25	99.12	98.88	10.20
ACP+RLM	78.59	76.71	74.06	240.38

Fuente: elaboración propia

Se generaron tres modelos diferentes para abordar la cuestión de multicolinealidad y comparar los estadísticos. Un modelo con RLM, otros dos donde se obtienen los componentes cuya variabilidad acumulada varia, y posteriormente se aplica regresión lo que corresponde a los modelos de ACP+RLM (1) y ACP+RLM (2).

El modelo RLM, que no incluyó la reducción de dimensionalidad, demostró ser altamente efectivo en la explicación de la variabilidad en los datos (Tabla 2).

Tabla 2 Resumen de estadísticos para modelos con multicolinealidad.

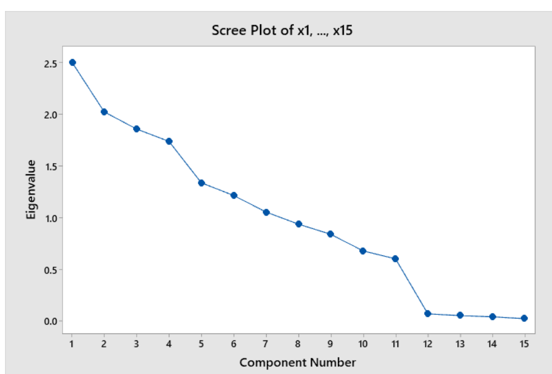
Modelo	R^2 (%)	R^2 ajustada (%)	R^2 de la predicción (%)	PRESS
RLM	99.40	99.30	99.15	10.07
ACP+RLM (1)	86.93	86.02	84.50	176.60
ACP+RLM (2)	97.32	97.04	96.61	40.65

Fuente: elaboración propia

Los resultados de este modelo revelaron un R^2 del 99.40%, lo que indica que el porcentaje de la variabilidad en la variable de respuesta que puede ser explicada por las variables predictoras. La R^2 ajustada y la R^2 de predicción también mostraron un rendimiento sobresaliente, con valores del 99.30 y 99.15%, respectivamente.

Además, el valor PRESS de 10.07 indica que el modelo es muy eficaz en la predicción de nuevos datos. En segundo modelo, ACP+RLM (1) (Tabla 2), donde se redujo la multicolinealidad para simplificar el modelo, se seleccionaron los componentes principales cuya variabilidad acumulada representara al menos el 85% del total de la variabilidad. En este caso, se utilizó en promedio de 8.63 componentes principales. A pesar de la reducción de dimensionalidad, este modelo logra explicar una cantidad considerable de la variabilidad en los datos, con un R^2 del 86.93%. Sin embargo, los valores de R^2 ajustada (86.02%) y R^2 de predicción (84.50%) son menores en comparación con el modelo RLM. El valor de PRESS aumentó considerablemente a 176.60 con respecto al modelo de RLM.

En tercer modelo, el cual cuenta con reducción de multicolinealidad, ACP+RLM (2), se obtuvieron los componentes principales mediante método gráfico. En la figura 1 se observa un descenso pronunciado entre el componente 10 y el 11, para este conjunto de datos en específico se tomaron 11 componentes para realizar la regresión lineal múltiple.

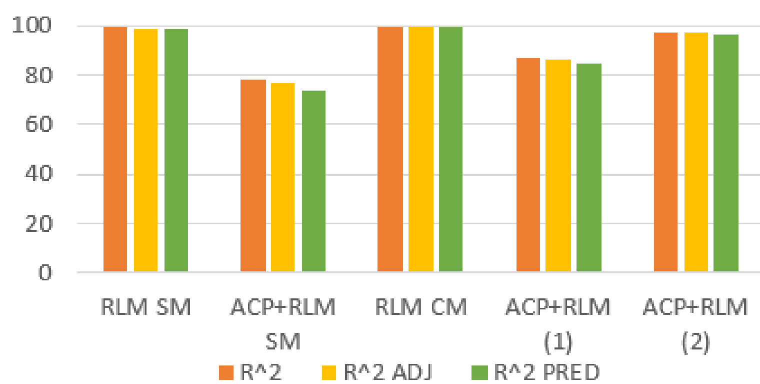


Fuente: elaboración propia

Figura 1 Gráfico de valores propios.

Se utilizó en promedio de 10.64 componentes para el tercer modelo. Estos explican del 90-95% de variabilidad acumulada. El R^2 (97.32%) fue alto y cercano al del modelo RLM, pero la R^2 ajustada (97.04%) y la R^2 de predicción (96.61%) fueron ligeramente inferiores. El valor de PRESS que fue de 40.65, aunque más bajo que en el modelo ACP+RLM (1) el cual es de 176.60, fue mayor que el del modelo RLM (10.07).

Para observar los efectos de la multicolinealidad en los modelos propuestos, se graficaron los estadísticos más relevantes (Figura 2). Se asignó SM, sin multicolinealidad, a ambos modelos que carecen de este efecto (RLM SM y ACP+RLM SM) y CM, con multicolinealidad, al modelo de regresión lineal múltiple que se ve influenciado mediante esta (RLM CM). Los modelos ACP+RLM (1) y (2) no sufren cambios en su denotación.



Fuente: elaboración propia

Figura 2 comparación de los modelos con efecto de la multicolinealidad.

Los modelos de RLM tanto CM como SM muestran valores altos para estos coeficientes. En cambio, los modelos de ACP+RLM muestran una diferencia significativa en los tres estadísticos cuando no hay multicolinealidad y cuando hay presencia de ella. El modelo ACP+RLM (2) es superior a los otros dos modelos teniendo valores de R^2, el R^2 ajustada y el R^2 predicha por encima de 95%, mientras que el modelo ACP+RLM SM tiene sus valores por debajo del 90%.

4. Discusión

Los resultados revelaron que el primer modelo no cuenta con multicolinealidad, es decir el modelo de RLM ya que sus valores del factor de inflación de varianza (VIF) se encontraban entre 1.03-1.41 [Chan et al., 2022]. Por otro lado, en el modelo ACP+RLM, la selección de componentes principales fue a partir de su variabilidad acumulada por encima del 85% [Jolliffe, 2013]. Este modelo tampoco cuenta con multicolinealidad ya que al momento de obtener los componentes y realizar la regresión lineal múltiple los valores VIF fueron de 1.0 [Chan et al., 2022].

Para estudiar el efecto de la multicolinealidad, la cual puede dificultar la interpretación de los modelos de regresión lineal múltiple, se simularon conjuntos de datos con multicolinealidad inducida. Los valores de VIF para ciertos coeficientes son mayores a 10 [Chan et al., 2022]. De acuerdo con los resultados para el modelo de RLM CM este efecto no representó un problema ya que los valores de los estadísticos están por encima del 99% y su PRESS no aumentó considerablemente lo que nos indica que el modelo puede predecir con precisión.

En los modelos donde se disminuye la multicolinealidad mediante ACP, se seleccionaron los componentes principales cuya variabilidad acumulada representara al menos el 85% del total [Jolliffe, 2013]. Para el modelo ACP+RLM (1) los valores de los estadísticos fueron mayores en comparación del modelo ACP+RLM SM. Esto sugiere que, aunque se redujo la multicolinealidad, la simplificación del modelo tuvo un impacto en su capacidad para explicar la variabilidad en los datos.

El tercer modelo, ACP+RLM (2), se basó en una selección gráfica de componentes principales en donde se observa una curvatura pronunciada dentro del gráfico de valores propios [Sanguansat, P. ,2012]. Aunque este enfoque permitió una reducción más eficaz de la multicolinealidad, los resultados fueron intermedios en comparación con los otros dos modelos.

Es evidente que el modelo de regresión múltiple con multicolinealidad (RLM CM) supera en la mayoría aspectos a los demás modelos, aunque el modelo RLM SM no es significativamente menor. Esto sugiere que la regresión lineal múltiple es eficiente para ambos tipos de conjuntos de datos, con y sin multicolinealidad. Por otro lado, los modelos que emplean el análisis de componentes principales seguido de regresión lineal múltiple (ACP+RLM) ofrecen una alternativa para abordar la multicolinealidad al reducir la dimensionalidad de los datos. Sin embargo, la simplificación del modelo mediante ACP puede tener un impacto negativo en la capacidad para explicar la variabilidad en los datos, como se observó en el modelo ACP+RLM (1). Esta simplificación puede llevar a una pérdida de información importante contenida en las variables originales, lo que afecta la capacidad predictiva del modelo.

Los resultados sugieren que, la aplicación del análisis de componentes principales (ACP) es más efectiva cuando los datos presentan multicolinealidad, dado que los estadísticos en los modelos ACP+RLM (1) y (2) presentan mejores resultados que del modelo ACP+RLM SM. Este enfoque podría proporcionar una estrategia valiosa para mejorar la robustez y la precisión de los modelos de regresión en presencia de multicolinealidad.

5. Conclusiones

Los resultados obtenidos revelan la importancia de abordar la multicolinealidad en el análisis de datos y la selección de modelos estadísticos. En ellos destacan la robustez y efectividad de la regresión lineal múltiple con un enfoque sólido para el análisis de datos y predicción de variables de respuesta. Para los conjuntos de datos que carecían de multicolinealidad, se observó que el modelo RLM obtuvo resultados sobresalientes, con un alto coeficiente de determinación (R^2) del 99.25%. Esto sugiere que el modelo RLM es altamente efectivo en la explicación de la variabilidad en la variable de respuesta. Además, tanto la R^2 ajustada como la R^2 de la predicción mostraron un gran rendimiento, respaldando la idoneidad de este enfoque.

En contraste, el modelo ACP+RLM en los conjuntos de datos sin multicolinealidad mostró resultados inferiores en términos de R^2 (78.59%) y en las métricas de R^2 de predicción y R^2 ajustada, aunque aún fueron significativos. El valor de PRESS fue notablemente mayor que en el modelo RLM, lo que indica una menor precisión en las predicciones.

Con respecto a la multicolinealidad en los conjuntos de datos, se generaron tres modelos distintos: RLM, ACP+RLM (1) y ACP+RLM (2). El modelo RLM, el cual no cuenta con reducción de dimensiones, explica la variabilidad en los datos con una R^2 del 99.40% y una alta precisión en las predicciones (PRESS de 10.07).

Los modelos ACP+RLM (1) y ACP+RLM (2) demostraron ser enfoques viables para abordar la multicolinealidad. Sin embargo, se observó que la simplificación del modelo en ACP+RLM (1) tuvo un impacto negativo en las métricas de rendimiento en comparación con el modelo RLM. Por otro lado, ACP+RLM (2) mostró un buen equilibrio entre reducción de multicolinealidad y capacidad de predicción, con resultados ligeramente inferiores a los del modelo RLM, pero aún significativos. Esto debido a que se consideró mayor variabilidad acumulada de los componentes a utilizar para la regresión.

El modelo RLM demostró ser altamente efectivo en la explicación de la variabilidad, por lo tanto, se recomienda el uso de la regresión lineal múltiple, especialmente en situaciones donde la multicolinealidad no sea un problema importante. Mientras que los modelos ACP+RLM ofrecen alternativas válidas para abordar la multicolinealidad, especialmente ACP+RLM (2), que equilibra reducción de dimensionalidad y capacidad predictiva.

Estos hallazgos son consistentes con investigaciones previas que han demostrado que el modelo de RLM es robusto ante la multicolinealidad y proporciona estimaciones precisas de los coeficientes de regresión. Adicionalmente, este estudio proporciona información valiosa para investigadores y profesionales que se enfrentan a problemas de multicolinealidad en el análisis de datos.

6. Bibliografía y Referencias

[1] Ahmed, N., Roy, S. R., & Islam, M. A. (2020). Forecasting Supply Chain Sporadic Demand Using Principal Component Analysis (PCA). https://api.semanticscholar.org/CorpusID:221680211

[2] Banakara, K. B., Pandya, H. R., & Garde, Y. A. (2019). Pre-harvest forecast of kharif rice yield using PCA and MLR technique in Navsari district of Gujarat. 21(3).

[3] Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics, 10(8), 1283. https://doi.org/10.3390/math10081283.

[4] Guillén-Casla, Vanesa & Rosales-Conrado, Noelia & León-González, María & Pérez-Arribas, Luis & Polo-Diez, LM. (2011). Principal component analysis

(PCA) and multiple linear regression (MLR) statistical tools to evaluate the effect of E-beam irradiation on ready-to-eat food. Journal of Food Composition and Analysis. 24. 456-464. 10.1016/j.jfca.2010.11.010.

[5] Gozá-León, O., Fernández-Águila, M., Rodríguez-Garcel, R. H., & Ojito-Magaz, E. (2020). Aplicación del Análisis de Componentes Principales en el proceso de purificación de un biofármaco. VacciMonitor, 29(1), 5-13.

[6] Harrell, F. E. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis (2da edición). Springer.

[7] Jolliffe, I. T. (2013). Principal Component Analysis. Springer New York. https://books.google.com.mx/books?id=-ongBwAAQBAJ

[8] Li, Q., Bessafi, M., & Li, P. (2023). Mapping Prediction of Surface Solar Radiation with Linear Regression Models: Case Study over Reunion Island. Atmosphere, 14(9), 1331. https://doi.org/10.3390/atmos14091331

[9] Minoglou, M. y Komilis, D. (2018). Describing health care waste generation rates using regression modeling and principal component analysis. Waste Management, 78, 811–818. https://doi.org/10.1016/j.wasman.2018.06.053

[10] Montgomery, D. C., Peck, E. A. y Vining, G. G. (2007). Introducción al análisis de regresión lineal (1a ed. en español, 3a ed. en inglés, 4a reimp). CECSA.

[11] Sanguansat, P. (2012). Principal Component Analysis Engineering Applications. InTech.

[12] Vervoort, S., & Wolff, M. (2020). Multivariate Spectra Analysis: PLSR vs. PCA + MLR. 7th International Electronic Conference on Sensors and Applications, 83. https://doi.org/10.3390/ecsa-7-08226.