Saber, Universidad de Oriente, Venezuela.Vol. 22 Nº 2: 181-184. (2010)

APLICACIÓN DEL ANÁLISIS FACTORIAL COMO UNA ALTERNATIVA DE SOLUCIÓN AL PROBLEMA DE MULTICOLINEALIDAD

APPLICATION OF FACTORIAL ANALISIS AS AN ALTERNATIVE FOR SOLVING THE PROBLEM OF MULTICOLINEARITY

Raúl Herrera León, Rosalvic Hernández Guevara

Departamento de Estadística, Universidad de Oriente, Núcleo de Nueva Esparta.
E-mail: raulherrera29@hotmail.com

RESUMEN

Se estudió la aplicación del análisis factorial en la solución del problema de multicolinealidad existente en un modelo obtenido por regresión lineal múltiple. Para ello, se utilizó la base de datos correspondiente a la variación de la calidad sensorial de la cerveza embotellada durante su almacenamiento a diferentes temperaturas. Mediante la regresión lineal múltiple se analizaron las variables respuestas (grado de oxidación, calidad de amargor y estabilidad organoléptica) como una función de la concentración de oxígeno de la cerveza, la temperatura y el tiempo de almacenamiento. Dado que todos los modelos encontrados presentaron multicolinealidad, se aplicó el análisis factorial para resolver este problema. Los modelos ajustados explicaron el 84,63%, 83,22% y 84,79% de la variabilidad en el grado de oxidación, calidad de amargor y estabilidad organoléptica respectivamente. Mediante la validación cruzada de estos modelos se encontraron altos valores en el coeficiente de determinación de predicción ($R^2 > 83\%$). Por lo tanto, el análisis factorial se puede utilizar para resolver el problema de la multicolinealidad en un modelo.

Palabras clave: Análisis de Regresión, Multicolinealidad, Análisis Factorial, Calidad Sensorial.

ABSTRACT

The application of factorial analysis as an alternative method to solve the multicollinearity problem existing in a model obtained by multiple linear regressions was studied. The data base used was the variation in the sensory quality of bottled beer during storage at different temperatures. Response variables (degree of oxidation, bitterness quality and organoleptic stability) were analyzed as a function of oxygen concentration, storage temperature and storage time by multiple linear regression. Since all of the obtained models showed multicollinearity, the factorial analysis was applied in order to solve this problem. The fitted models explained 84.63%, 83.22% and 84.79% of the variability in the degree of oxidation, bitterness quality and organoleptic stability, respectively. High prediction of the coefficients of determination ($R^2 > 83\%$) for these models were found by application of crossed validation. Therefore, the factorial analysis can be used to solve the multicollinearity problem in a model.

Key words: Regression analysis, Multicollinearity, Factor Analysis, Sensory Quality.

INTRODUCCIÓN

El análisis de regresión lineal múltiple es una técnica estadística que se utiliza para analizar la relación existente entre una variable dependiente o respuesta (Y) y un conjunto de otras variables denominadas regresoras, predictoras o explicativas (X_1, X_2, …, X_k), cuyo primer propósito es, medir el efecto relativo de cada variable explicativa sobre la respuesta (Y) y el segundo, prever el valor de la variable respuesta una vez conocido el valor de las variables explicativas.

Sin embargo, para que se pueda medir el efecto preciso de cada variable explicativa en la dependiente, es imprescindible la ausencia de multicolinealidad; es decir, que no haya correlación entre las variables predictoras incluidas en el modelo de regresión. Dado que la existencia de correlación elevada entre dos o más variables (multicolinealidad) repercute, de manera directa, en los errores estandares de los estimadores mínimos cuadrados ordinarios de los parámetros relacionados con dichas variables; éstos se ven indebidamente incrementados, lo que provoca que la estimación de los coeficientes sea menos precisa (coeficientes inestables y con gran varianza). Además, el cálculo de la ecuación de regresión, producto de que la misma requiere la inversión de la matriz $\mathbf{X'X}$ y si una de las variables explicativas es combinación lineal exacta de las demás, la matriz que contiene los valores de las variables regresoras ($\mathbf{X}$) tendrá un rango menor que el número de parámetros a estimar, razón por lo cual la matriz $\mathbf{X'X}$ será singular y el sistema de ecuaciones no tendrá solución única.

Al respecto, se pueden aplicar diferentes métodos

Recibido: noviembre 2008. Aprobado: junio 2010.
Versión final: junio 2010

para evitar los efectos producidos por la presencia de multicolinealidad en un análisis de regresión, siendo el método de estimación sesgada (Regresión por Componentes Principales y Regresión Ridge) uno de los más utilizado. Aún cuando, Peña (2002), señala que existen dos únicas soluciones, la primera de ellas es "eliminar regresoras", reduciendo el número de parámetros a estimar y la segunda "incluir información externa a los datos". La solución más simple es eliminar de la ecuación variables altamente correlacionadas con otras, pero ésta es la opción más drástica, pues provoca una amplia disparidad de incógnitas en relación a: ¿Cuál o cuáles variable(s) predictora(s) debe(n) omitir(se)?.

Para dar respuesta a esta interrogante, Cea (2004) recomienda la aplicación del análisis factorial previa al análisis de regresión lineal para la identificación de variables explicativas colineales a ser excluidas. Esto permite obtener un modelo parsimonioso; es decir, con un número pequeño de parámetros, donde sus coeficientes estimados cumplen con las propiedades de un buen estimador (insesgabilidad, consistencia, suficiencia y eficiencia) y a su vez incluye sólo aquellas variables predictoras que muestren ser relevantes en la predicción de la variable dependiente. Razón por la cual, no es aconsejable la inclusión de muchas variables regresoras en el análisis de regresión, a menos que se demuestre que son relevantes para la predicción de la variable dependiente.

En este sentido, para ilustrar la aplicación del análisis factorial en la resolución del problema de multicolinealidad de manera empírica, se utilizó para ello una base de datos extraída del trabajo de grado intitulado "Predicción sesgada de la calidad sensorial de la cerveza de lata y botella en función de las mediciones instrumentales y de los factores que la afectan" realizado por Rodríguez y Rodríguez (2001), donde se presentan modelos de predicción sesgada para tres variables dependientes: grado de oxidación, calidad de amargor y estabilidad organoléptica, especialmente para la cerveza embotellada, en función de siete variables predictoras: concentración de oxígeno, temperatura, tiempo de almacenamiento, contenido de dióxido de carbono, ácidez iónica, amargor y concentración de ácido tio-barbitúrico; dado que estas variables explicativas están correlacionadas.

MATERIALES Y MÉTODOS

Las variables utilizadas para establecer la relación funcional entre la calidad sensorial de la cerveza de botella y las características fisicoquímicas y los factores que la afectan, obtenidas de una industria cervecera del estado Anzoátegui, fueron codificadas de la siguiente manera:

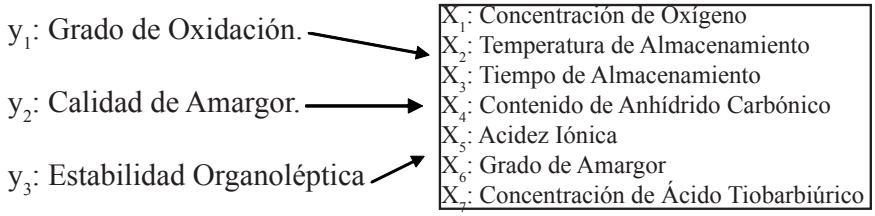

Figura.1 Codificación de las variables dependientes y explicativas

El conjunto de datos está constituido por 768 muestras que se dividieron en dos lotes iguales para las concentraciones de oxígeno, a su vez las 384 muestras fueron subdivididas en lotes iguales para almacenarlas a las tres temperaturas (5, 28 y 60^0C) y luego cada 7 días por espacio de 16 semanas, se llevaron a cabo 8 mediciones instrumentales del contenido de dióxido de carbono, acidez iónica, concentración de ácido tiobarbitúrico y amargor; además de 8 mediciones sensoriales del grado de oxidación, la calidad de amargor y la estabilidad organoléptica, con ayuda de un panel entrenado.

Esta base de datos, se dividió en dos muestras de igual tamaño, la primera para estimar los modelos de regresión lineal múltiple sobre las características sensoriales (grado de oxidación, calidad de amargor y estabilidad organoléptica) en función de las características fisicoquímicas (dióxido de carbono, acidez iónica, amargor, concentración de ácido tiobarbitúrico) y de los factores que las afectan (oxígeno, temperatura y tiempo de almacenamiento), dado que éstas variables explicativas están correlacionadas unas con otras, es decir, existe problema de multicolinealidad en los datos, se procedió a realizar un análisis factorial exploratorio como una técnica alternativa que permite agrupar variables independientes muy correlacionadas entre sí en un número menor de factores extraídos, representados por las variables explicativas cuyas cargas factoriales sean significativas y de mayor valor. Estas variables fueron utilizadas para aplicar el análisis de regresión múltiple,

con el cual se estableció la relación existente entre dichas variables y la calidad sensorial de la cerveza embotellada. La segunda muestra fue empleada para validar los modelos obtenidos, que luego se compararon con los modelos arrojados mediante los métodos de estimación sesgada Regresión Componentes Principales y Regresión Ridge reseñados por Rodríguez y Rodríguez (2001).

En la aplicación del análisis factorial exploratorio y el posterior análisis del modelo de regresión múltiple para estimar y/o predecir la calidad sensorial de la cerveza embotellada, se utilizaron los paquetes de cómputo estadísticos Statgraphics Plus bajo Windows Versión 5.1 y SPSS bajo Windows Versión 10.0; mientras que para la validación de los modelos obtenidos, se empleó la hoja de cálculo electrónica Microsoft Excel.

RESULTADOS Y DISCUSIÓN

Para detectar la existencia de multicolinealidad se procede, al análisis de la matriz de correlaciones que figura en la Tabla 1. La inspección de esta matriz revela que 16 de las 21 correlaciones (76%) son significativas al nivel de significancia del 1%, destacándose correlaciones elevadas entre las variables: X_2 y X_4 (-0,713); X_2 y X_7 (0,798); X_3 y X_5 (0,765); X_4 y X_7 (-0,953) y X_5 y X_6 (-0,763). Siendo esto un indicio de posible presencia de multicolinealidad en los datos. Dicha matriz también presenta un determinante muy cercano a cero (|R|=0,0008995) y los valores de la diagonal principal de su inversa (Factores Infladores de la Varianza) se sitúan por encima del valor de referencia cinco (5), lo que corrobora lo señalado con relación en la existencia de intercorrelaciones muy elevadas entre las variables explicativas.

Tabla 1. Matriz de correlaciones para el conjunto de variables explicativas.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1,000	0,000 (1,000)	0,000 (1,000)	0,022 (0,666)	0,284 (0,000)	-0,300 (0,000)	0,050 (0,325)
X_2		1,000	0,000 (1,000)	-0,713 (0,000)	0,355 (0,000)	-0,551 (0,000)	0,798 (0,000)
X_3			1,000	-0,431 (1,000)	0,765 (0,000)	-0,587 (0,000)	0,308 (0,000)
X_4				1,000	-0,609 (0,000)	0,622 (0,000)	-0,953 (0,000)
X_5					1,000	-0,763 (0,000)	0,571 (0,000)
X_6						1,000	-0,615 (0,000)
X_7							1,000

Las cifras entre paréntesis corresponden a la significatividad de las correlaciones bivariables respectivas

Las cifras entre paréntesis corresponden a la significatividad de las correlaciones bivariables respectivas.

De la interpretación del determinante de la matriz anterior (Tabla 1) también puede afirmarse la presencia de varianza común de las variables, ayudando a la agrupación u obtención de combinaciones lineales de variables correlacionadas. Lo cual constituye la comprobación de la pertinencia de la aplicabilidad del Análisis Factorial Exploratorio, como una solución alternativa al problema de multicolinealidad, evitando los efectos ocasionados por esta al estimar un modelo de regresión. Por consiguiente, se tiene la aplicación del análisis factorial a la base de datos analizada, empleado el método de extracción de factores componentes principales y como criterios a seguir en la determinación del número de factores: los autovalores y el porcentaje de varianza total atribuible a cada factor. No obstante, basándose en el patrón de altas cargas factoriales la interpretación resulta bastante difícil, por esta razón se recurrió al método de rotación varimax para redistribuir la varianza del primer factor al segundo factor, obteniéndose los siguientes resultados:

Tabla 2. Matriz factorial rotada por el procedimiento ortogonal varimax.

Variables	Factores		Comunalidades
	1	2	
X_2	0,945*	-0,001	0,892
X_7	0,917*	0,302	0,931
X_4	-0,846*	-0,406	0,880
X_3	-0,003	0,955*	0,911
X_5	0,346	0,874*	0,882
X_6	-0,531	-0,690*	0,759
			Total
Suma de cuadrados (Autovalores)	2,850	2,410	5,260
% varianza total	47,49	40,10	87,59

* Factor en el que más satura las variables

De la matriz factorial rotada (Tabla 2), se seleccionó como representante de cada factor la variable con mayor coeficiente factorial en dicho factor. En este sentido, la variable predictora X_2 representa al factor 1, mientras que el factor 2 esta representado por la variable predictora X_3. Razón por la cual, se procedió a realizar el análisis de regresión lineal múltiple para estimar la calidad sensorial de la cerveza embotellada en función de las variables explicativas seleccionadas y la variable predictora (X_1), pues ella por sí sola representa un factor. Cabe resaltar, que

las variables explicativas X_1, X_2 y X_3 no están correlacionadas, esto se evidencia observando la matriz de correlaciones presentada en la tabla 1, donde las correlaciones entre las tres (3) variables son iguales a 0. Por ende, los modelos construidos son:

$$\hat{y}_3 = 4{,}2920 - 1{,}0789X_1 - 0{,}0404X_2 - 0{,}0801X_3$$

$$\hat{y}_3 = 4{,}2920 - 1{,}0789X_1 - 0{,}0404X_2 - 0{,}0801X_3$$

$$\hat{y}_3 = 4{,}2920 - 1{,}0789X_1 - 0{,}0404X_2 - 0{,}0801X_3$$

Tabla 3. Valores de los coeficientes de determinación ajustado $R^2_{aj}(\%)$ y de predicción $R^2_{pd}(\%)$ para los modelos de las características sensoriales de la cerveza embotellada, obtenidos mediante los métodos de estimación sesgado y el método de mínimos cuadrados ordinarios.

	Variables Dependientes					
	Grado de Oxidación		Calidad de Amargor		Estabilidad Organoléptica	
	Estimación	Predicción	Estimación	Predicción	Estimación	Predicción
Métodos de Estimación Aplicado	$R^2_{aj}(\%)$	$R^2_{pd}(\%)$	$R^2_{aj}(\%)$	$R^2_{pd}(\%)$	$R^2_{aj}(\%)$	$R^2_{pd}(\%)$
Regresión por Componentes Principales	85,82	85,76	84,84	85,55	85,79	85,52
Regresión Ridge	83,34	87,45	82,05	87,45	83,10	87,36
M.C.O. con las variables seleccionadas mediante el análisis factorial	84,63	84,95	83,09	84,64	84,79	84,74

CONCLUSIONES

El análisis factorial presenta un método alternativo para la solución del problema de multicolinealidad presente en un modelo de regresión, pues proporciona una compresión clara de cuáles de las variables explicativas están correlacionadas, razón por la cual se puede seleccionar, de cada factor estructurado, una variable que lo represente. Estas variables suelen coincidir con aquellas de mayor coeficiente factorial en dicho factor y son las que se deben considerar para estructurar los modelos.

Los modelos de regresión múltiple de las características sensoriales de la cerveza embotellada producto de la previa aplicación del análisis factorial, presentan coeficientes de determinación ajustados y de predicción muy similares a los modelos arrojados por los métodos de estimación sesgada Regresión por Componentes Principales y Regresión Ridge (Tabla 3). Sin embargo, los modelos obtenidos presentan un menor número de parámetros, cumpliendo así con el principio parsimonioso (escoger el modelo más sencillo) y con coeficientes estimados que cumplen con las propiedades de un buen estimador (insesgabilidad, consistencia, suficiencia y eficiencia), dado que se obtuvieron mediante el método de mínimos cuadrados ordinarios.

REFERENCIAS BIBLIOGRÁFICAS

Cea, M. 2004. Análisis Multivariable: Teoría y Práctica en la Investigación Social. Editorial Síntesis, C.A. (2da. Edición). Madrid, España. pp. 14-58, 428-504.

Rodríguez C., Rodríguez R. 2001. "Predicción sesgada de la calidad sensorial de la cerveza de lata y botella en función de las mediciones instrumentales y de los factores que la afectan". Trabajo de Grado no publicado. Universidad de Oriente, Núcleo de Nueva Esparta.

Peña, D. 2002. Regresión y Diseño de Experimentos. Editorial Alianza. México, D.F. pp. 142-162.