# Técnicas estadísticas multivariantes para valorar la satisfacción de clientes

**Autor:** Ángel Gómez Degraves, Ph.D. Tecana American University

gomezdegraves@gmail.com

Publicación: junio 21, 2021

Una de las personas que escribe sobre las técnicas de análisis multivariante (TEM), Figueras (2000) define el análisis multivariante como el conjunto de métodos estadísticos que tienen por objeto analizar simultáneamente conjuntos de datos multivariantes donde existe varias variables que se miden en un solo individuo, objeto o concepto. El autor define al análisis multivariante como el conjunto de técnicas matemáticas – estadísticas, que permiten explicar y predecir variables aleatorias en función de otras de diversa índole y resumir, explicar e interpretar la estructura de varianza o de correlación de los datos, con el fin de generar conocimiento oculto que permita la toma de decisiones.

Algo esencial de las TEM es comprender los supuestos teóricos necesarios en la aplicación de una determinada técnica. En este sentido, hay varios aspectos a considerar: el objetivo del estudio, el tipo de datos: cualitativo o no métrico y cuantitativo o métrico. La escala de medición no métricas (nominales o de categoría y ordinales) y métricas (de intervalo o razón); el tipo de distribución de probabilidades de las variables, el número de variables y el número de individuos u objeto de estudio.

Lo anterior, permite deducir lo complejo del Análisis Multivariante y el cuidado que se debe tener al seleccionar una TEM cuando se procede a analizar un conjunto de datos.

Existen varias formas de abordar las TEM, una de ellas es mediante el enfoque de dependencia o Inferencial y el de interdependencia o descriptivo.

El análisis de dependencia es el que resulta más familiar y las técnicas en este enfoque se consideran como una extensión del análisis estadístico univariado y bivariado.

El análisis de dependencias es aquel en el que una variable o conjunto de variables se consideran como dependiente y su variabilidad va a ser explicada por una o

más variables independientes. El investigador hace varios supuestos en relación a los datos y en base a ellos, selecciona la técnica adecuada para el análisis de los datos.
En el análisis de dependencia, el objetivo es determinar si el conjunto de variables independientes afecta y en qué forma, al conjunto de variables dependientes; es decir, si el conjunto de variables independientes afecta al conjunto de variables dependientes, individualmente o conjuntamente.

En cambio, el análisis de interdependencia o descriptivo trata de analizar cómo y cuales variables se relacionan entre sí, no hay restricciones en cuanto a hipótesis estadísticas como en el caso del análisis de dependencia. No se plantean hipótesis previas y se resume la formación para describir el comportamiento de un fenómeno con el menor número de variables en el análisis de interdependencia persiguen identificar y entender la estructura de correlación o covariación de los datos, utilizando la matriz de varianzas – covarianzas o en su lugar, la matriz de correlaciones, su principal característica es la reducción de la dimensión de los datos, logrando que el problema se observe en una forma más simple.

Las TEM en el cual el análisis de interdependencia no son una continuación de las univariadas y bivariadas y no se requieren los supuestos en los datos y variables que requieren las del análisis de dependencia, además no existe la presencia de variables independientes y dependiente.

**Regresión Lineal Múltiple**

Esta técnica estadística multivariante (TEM) se utiliza cuando se analiza la relación entre una variable dependiente cuantitativa y varias variables dependientes o explicativas predictoras que pueden ser combinaciones de variables cuantitativas (métricas) o cualitativas (categóricas o no métricas).

La regresión lineal múltiple (RLM), busca una función de regresión poblacional partiendo de una función de regresión muestral. En su esencia el modelo de RLM une aspectos matemáticos, teóricos de la disciplina tratada y estadísticos, con el fin de establecer predicciones, influencias de variables independientes y detectar interacciones significativas entre las variables explicativas o predictoras que afecten la variable respuesta o dependiente. Para utilizarla se debe tener en cuenta el conocimiento y

experiencia del fenómeno que se estudia, no es solo la aplicación de la técnica como único fin.

En los modelos de RLM, las variables explicativas o independientes pueden ser continuas (métricas) o cualitativas. Hay casos donde se tienen combinaciones de variables, dicotómicas, politómicas y numéricas; es decir, se pueden tener variables cualitativas como cuantitativas como predictoras. En el caso de que se tenga variables cualitativas o categóricas explicativas, se introduce nuevas variables denominadas dummy o ficticias, donde en el caso de una variable con 2 categorías o dicotómica, se codifica como 0 y 1.

En el caso que la variable categórica tenga más de 2 categorías o modalidades se utilizan K-1 variables indicadoras o dummy, donde K es el número de categorías, y se codifican con 0 o 1, teniendo una base como patrón de comparación, por ejemplo, si hay 3 categorías se puede usar 2 variables indicadoras. Por otro lado, las técnicas conocidas como análisis de la varianza (ANOVA), se consideran como un caso especial de RLM, donde las variables independientes son cualitativas o categóricas y se representan como dummy o falsa variable. Además, si se incluye una o más variables independientes cuantitativas (continuas discretas) aparece el análisis de la covarianza (ANCOVA), el cual puede ser simple, si hay una sola variable cuantitativa explicativa, además de las dummy del ANOVA, y si son varias covariables numéricas, al modelo se le llama análisis de la covarianza múltiple.

En la RLM existe lo que se conoce como multicolinealidad, la misma, se refiere solo a las relaciones lineales entre las variables explicativas, no a las relaciones no lineales. La RLM al igual que todas las técnicas estadísticas multivariantes requiere que el número de observaciones o elementos de análisis (n) exceda el número de variables independientes o predictoras.

La RLM es muy útil en estudios de satisfacción de clientes, sobre todo en encuestas de satisfacción de clientes, donde se utiliza el puntaje global de satisfacción como variable dependiente o respuesta y las dimensiones de ese constructo como variables independientes, para identificar las dimensiones de mayor y menor impacto en la satisfacción, también se puede utilizar el puntaje de cada dimensión de satisfacción como variable dependiente y los indicadores de cada dimensión serían las variables explicativas o independientes. Así mismo, Para predecir o pronosticar valores de indicadores determinantes del negocio (volumen de venta entre otros) basándose en su relación con otras variables explicativas o independientes.

Por otro lado, la RLM se usa frecuentemente en combinación con el análisis de Componentes Principales, del cual se extraen los primeros componentes que explican la mayor parte de la varianza total y esas nuevas variables no correlacionadas se utilizan como variables independientes en la RLM. Esto evita la multicolinealidad en las variables independientes. Este análisis se utiliza mucho cuando las variables independientes o predictoras están altamente correlacionadas.

**Regresión Logística**

En las ciencias naturales y sociales es muy común la aplicación de la Regresión Logística (RL) como herramienta de análisis estadístico multivariante, como lo menciona Alderete (2006). El modelo RL es apropiado para predecir la probabilidad de tener una respuesta particular en función de un conjunto de variables predictoras, explicativas o independientes, que generalmente se consideran factores de riesgo, se utiliza mucho en los fenómenos de las ciencias de la salud. La variable dependiente a predecir es categórica (dicotómica o politómica).

La RL produce una ecuación, donde se relaciona la probabilidad de un resultado para el valor particular de dos (2) o más variables independientes:

$$P\ (Y=1) = \frac{1}{1 + e^{-\left(\beta_0 - \beta_1 X_1 \mid \beta_2 X_2 - \cdots - \beta_k X_k\right)}}$$

$$P\ (Y=1) = \frac{e^{\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k\right)}}{1 + e^{\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k\right)}}$$

donde:

P (Y=1) = Probabilidad de que la variable aleatoria Y tome el valor uno (1=éxito) en presencia de covariables o factores de riesgo.

$\beta_0$ =constante del modelo.

K = número de variables independientes.

$\beta_i$ = coeficiente de Regresión Logística.

$X_i$ = covariable o variable independiente.

$e$ = base del logaritmo natural (2,718).

El modelo de RL tiene la ventaja de que no requiere el supuesto de normalidad multivariante y de homocedasticidad; además, puede incorporar efectos no lineales y variables dummy y cuantitativas discretas y continúas. Al no cumplirse los supuestos del Análisis Discriminante (AD), se utiliza la RL.

En relación a la estimación de parámetros del modelo, utiliza el método de máxima verosimilitud en lugar de mínimos cuadrados, para el ajuste del modelo se utiliza el valor -2LL o dos (2) veces el logaritmo de la razón de verosimilitud. Si el valor de -2LL es pequeño se considera un ajuste adecuado del modelo. Hay varios indicadores de bondad de ajuste del modelo, uno de ellos muy útil es el de Hosmer y Lebeshow (citado por Alderete, 2006). Para la evaluación global del modelo de RL se utiliza el estadístico ERV= $\{-2\ln[\mathbf{L}(\mathbf{R})]\} - \{-2\ln[\mathbf{L}(\mathbf{MC})]\}$ . R es el modelo reducido y MC es el modelo completo. Este estadístico sigue una distribución ji-cuadrada con k-1 grados de libertad y k es el número de parámetros incluidos en el modelo completo, que han sido estimados por máxima verosimilitud.

Para contrastar los coeficientes (β's) de RL, se utiliza el estadístico de Wald, con el cual se verifica la hipótesis estadística $H_o$: β =0 y $H_1$: β ≠ 0, donde W=$\beta_j/s_{j,,}$ donde β y s son estimadores, y W tiene una distribución Ji-cuadrada con 1 grado de libertad. Si se rechaza $H_o$ ($p<0.05$) entonces $X_j$ afecta la probabilidad de éxito en la RL. Aquella variable con valores bajos de W, cerca de 0, se excluye del modelo.

Para la satisfacción de clientes, se utiliza la RL para la estimación y predicción de la probabilidad de estar completamente satisfecho con el servicio que ofrece la empresa y para identificar aquellas dimensiones e indicador o variables explicativas que tiene mayor importancia en el nivel de satisfacción de los clientes. Es decir, determinar la influencia de las variables independientes en la probabilidad de aparición o ocurrencia del evento P (Y=1), done 1= éxito y 0= falla. Predecir la respuesta P (Y=1) de presencia del evento mediante la ecuación logística en cada sujeto, se puede predecir si un cliente potencial podrá ser un cliente satisfecho o no, con el servicio. Clasificar el sujeto o individuo según la probabilidad que tenga de pertenecer a determinado grupo (1=satisfecho y 0= insatisfecho), si P (Y=1)>=0,5 y P (Y=1) <0,5 y calcular el peso para cada variable independiente, mediante el Odds Ratio (OR): $e^b$ , donde b es el coeficiente de Regresión

logística de la variable independiente en cuestión. Si OR>1 se considera factor de peso, si OR=1 se considera un factor de poco o ningún peso.

**Análisis Discriminante**

El Análisis Discriminante (AD) es una técnica estadística multivariante de dependencia que tiene dos fines: descriptivo y predictivo. El descriptivo es determinar cuál de las variables explicativas o independientes son las que diferencian los grupos formados a priori, cuales son importantes para clasificar los sujetos u objetos en dos (2) o más clases. El fin predictivo es proporcionar una regla de clasificación mediante la cual se determina a qué grupo pertenece una observación o sujeto. Se predice la pertenencia de un sujeto a un grupo. Se calcula la probabilidad de que un caso sea incluido en un grupo, basada en valores de variables independientes.

En el AD se considera la variable dependiente como categórica (dicotómica o politómica) y las variables independientes son cuantitativas (métricas) en escala de intervalo o razón. En RLM se explica o predice una variable dependiente cuantitativa en función de variables cuantitativas o cualitativas.

El AD funciona mejor si las variables explicativas tienen distribución normal multivariada, usándose generalmente variables estandarizadas. Las variables que se utilizan para discriminar o separar grupos no deben ser redundantes o no debe existir multicolinealidad entre ellas. La relación entre la variable dependiente e independientes debe ser lineal. Si no se cumplen los supuestos y la variable dependiente es categórica, se utiliza la RL. La RL es más flexible en relación con los supuestos. Generalmente LA RL se aplica al caso de dos (2) grupos y AD para más de dos grupos. Las variables explicativas del AD y de la RL se seleccionan de la teoría, investigaciones previas y de la experiencia del investigador. En el AD se tienen clasificados los individuos en grupo a priori, en los que se midieron las variables cuantitativas independientes.

Antes de proceder a realizar un AD se realiza una prueba de hipótesis estadística sobre los vectores de medias para los diferentes grupos formados, si no se encuentra diferencia significativa ($p<0.05$) entre los vectores de medias, no se realiza el AD.

Las técnicas de AD se clasifican por el número de grupos o clases de la variable dependiente. Si tiene dos (2) niveles se conoce como discriminante de Fisher y si tiene más de dos grupos se conoce como múltiple. Para el caso de dos (2) grupos se obtiene un función discriminante y para más de dos (2) grupos se obtiene g-1 funciones, el número de grupos menos uno.

La función discriminante se define como una combinación lineal de las variables independientes del tipo:

$D = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$

donde:

D = puntaje discriminante.

$\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ son los coeficientes discriminantes.

$X_1, X_2, X_3, \ldots, X_k$ son las variables independientes.

Para estimar los coeficientes de la función discriminante hay dos procedimientos: el directo y el de pasos (stepwise). En el directo se incluyen todas las variables en el modelo y en el de pasos el investigador selecciona un subconjunto de variables explicativas.

El AD como técnica clasificatoria puede utilizarse también, después de un Análisis de Conglomerados (AC) o Clúster Analysis para chequear dicha clasificación, donde con una medida de probabilidad se asigna los individuos a los grupos y se verifica la clasificación anterior.

En la satisfacción de clientes, se utiliza el AD en el Desarrollo de la(s) funcione(s) discriminante(s) que son combinaciones lineales de las variables independientes que discriminen los grupos (dos o más), dependiendo de cuantos grupos de satisfacción de clientes se formen, de acuerdo a los valores que resulten de la encuesta. Se Identificarán las variables de predicción o independientes que contribuyen a explicar la diferencia entre los grupos, se establecerá un procedimiento o regla para clasificar individuos en base a los valores de las variables explicativas. Se creará un modelo de AD que permite clasificar cualquier cliente potencial, en cuanto a su grado de satisfacción con el servicio que ofrece la empresa en estudio.

**Análisis de Factores Exploratorio**

Es una técnica estadística de interdependencia entre variables utilizadas para explicar las interrelaciones entre un número relativamente elevado de variables métricas observables en términos de un número menor de variables observables llamados factores o variables ocultas; esto es, los factores no son medibles directamente, están subyacentes (Figueras, 2000; Schuschny y Soto 2009; Richarm, 1997; Zikmund, 1998), las variables observadas se modelan, como combinaciones lineales de Factores Comunes más un término específico llamado error, conocido como la especificidad de la variable.

Se trata de explicar lo común que tiene una variable con otras en términos de los Factores Comunes; es decir, la varianza común entre las variables, siendo la varianza común aquella parte de la varianza total de una variable que es compartida con las demás variables originales. Lo que ocurre es una reducción en la dimensión del problema, se parte de un gran número de variables y se llega a reducir ese número en unas variables no observables llamadas factores o dimensiones, que explican gran parte de lo que podría explicar el conjunto original de variables.

El modelo matemático del AF (Chatfield y Collins, 1980; Gondar, 2000), supone que cada variable observada es función de un número de factores comunes y un factor específico:

$$X_j = m_{j1}f_1 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + m_{jm}f_m + e_j$$

. .

. . j=1,2,…., p Variables

. . k=1,2…..,m Factores

. .

$$X_p = m_{p1}f_1 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + m_{pm}f_m + e_p$$

donde:

$\{m_{jk}\}$= Importancia del factor k sobre la j-ésima variable.

$m_{jk}$ es llamada la carga o saturación de la variable j en el factor k.

$f_i$ es llamado factor común.

$e_j$ = es el factor específico.

Se supone que los factores comunes o variables latentes tienen media 0 y varianza 1. Los factores específicos o únicos son variables con media cero y las varianzas pueden ser distintas y están no correlacionados entre sí. Así mismo, los factores comunes y específicos no se correlacionan. Más que una TEM se refiere a un conjunto de técnicas donde se representa un conjunto de variables originales observables a través de un conjunto de variables no observables o hipotéticas.

En otro orden de ideas, el AF explica la estructura de correlaciones o covarianza de las variables originales en función de nuevas variables, llamadas factores latentes u ocultos. Los factores comunes representan lo común entre las variables originales y que explica la diferencia entre los sujetos u objetos de análisis.

Pocos factores comunes son los responsables de la correlación o covariación entre las variables originales y contienen la mayor parte de la información.

Las variables se miden en una escala de intervalo o razón (métrica). El AF crea un conjunto de variables no correlacionadas llamadas factores a partir de variables correlacionadas, si las variables originales no están correlacionadas, no hay razón para realizar un AF, no hay nada que explicar de las correlaciones de las variables originales (Johnson, 2000; Gondar, 2000).

En una investigación de satisfacción de clientes, se utiliza el AF exploratorio con el fin de identificar los factores hipotéticos o dimensiones que subyacen en la satisfacción de los clientes, relacionando esos factores comunes con las variables, ello lleva a comprender más el fenómeno de satisfacción de los clientes como un aspecto básico de la calidad del servicio ofrecido por la empresa. Así mismo, el AF permite seleccionar los factores que explican la diferencia entre los clientes, para manejarlos mediante la fijación de estrategias de mejora del grado de satisfacción de los clientes. En una muestra piloto o pre muestra se puede usar el AF para eliminar ítems en el cuestionario

de la encuesta. Eliminándose preguntas que no aportan variabilidad en la muestra, cosa que también permite el AF.

**Análisis de Conglomerados**

Otra de las técnicas estadísticas multivariantes que se aplicará en este estudio, es el Análisis de Conglomerados (AC), este es un conjunto de técnicas que permiten clasificar una muestra de objetos, casos, variables (entidades) en grupos relativamente homogéneos llamados Conglomerados (Clústeres), donde los objetos de cada grupo tienden a ser similares entre sí y diferentes a los objetos de otros grupos (Malhotra, 1998; Gondar, 2000; Figueras, 2001). Al Análisis de Conglomerados se le conoce como taxonomía numérica o reconocimiento de patrones (Miquel et al., 1997) y se diferencia del AD, en que se desconoce el número de grupos y su composición; es decir, los grupos no se conocen a priori, sino que se forman post-hoc.

Se realiza la división de los objetos de tal manera, que los perfiles de ellos en un mismo grupo sean muy similares entre sí y muy diferentes al compararlos con los otros grupos.

Para la aplicación del AC no se requiere de supuestos estadísticos en cuanto a las distribuciones de probabilidades de las variables originales y estas pueden ser métricas y no métricas (ordinales y binarias, incluso) y al igual que las otras TEM, el número de individuos debe ser mayor que el número de variables.

En cuanto al tipo de variables, si son cuantitativas, se utilizan las distancias o similaridades y antes de su cálculo, se procede a estandarizar las variables originales a fin de que las diferencias entre las unidades de medida no afecten las distancias o similaridades, si las variables son cualitativas, se pueden identificar con los valores cero (0) o uno (1) en dos categorías y se pueden utilizar también variables en escala ordinal.

En el AC, hay si se quiere, dos aspectos clave para el análisis de los datos, una medida de distancia o similaridad y el método de formación de los grupos o Clústeres. Los datos para realizar la conglomeración o tipificación pueden ser:

- Una matriz de distancia o similaridades donde las hileras y columnas corresponden a objetos.
- Una matriz de datos donde los objetos son las hileras y las variables las columnas.

Un aspecto importante es la selección de las variables para realizar el AC, hay que tener en cuenta la teoría, investigaciones previas y experiencia del investigador sobre el tema de satisfacción de clientes, es importante aquí, hacer ver que quien formará los grupos o clústeres finales será el investigador, apoyado por la técnica de AC y su experiencia.

Una de las características que tienen las técnicas estadísticas multivariantes, incluyendo el AC, es que se pueden combinar con otras técnicas para la solución de problemas complejos, donde hay un gran volumen de datos. Así por ejemplo, se puede utilizar un AF y luego un AC con los componentes importantes, esto reduce la complejidad de la clasificación de los objetos.

Se puede utilizar un AC y después de constituidos los conglomerados, se puede usar un AD con el fin de ayudar a describir los perfiles de grupos e identificar las variables con mayor peso en la formación de los grupos.

Si se tiene los resultados de un AC, donde se ha definido los grupos o Clústeres, se puede usar un AD para asignar un nuevo individuo a los grupos de acuerdo a funciones discriminantes y con los puntajes de cada individuo en la función discriminante, se utiliza un Análisis de la Varianza (ANOVA) para evaluar si hay diferencias entre las medias de grupo.

Para realizar un AC hay que demostrar que existen fuertes correlaciones en las variables que van a constituir el perfil de cada conglomerado. Si no existe fuerte correlación entre las variables, no tiene sentido realizar un AC.

Para la investigación de satisfacción de clientes, se utiliza una matriz de datos provenientes de la encuesta de satisfacción de clientes, sobre las variables que el equipo investigador considere para realizar una segmentación de clientes, que va desde clientes insatisfechos completamente hasta clientes satisfechos completamente. Los conglomerados serán de clientes. Esta segmentación de clientes proporciona el perfil de cada grupo formado y se identificarán las variables que discriminan los grupos mediante un AD. Se presenta una caracterización completa de los grupos de satisfacción que se formen en la segmentación.

**Referencias bibliográficas**


Alderéte, A. (2006). Fundamentos del Análisis de la Regresión Logística en la Investigación Psicológica. *Evalua, 6,* 52-67.

Chatfield, C. y Collins, A. (1980). *Multivariate analysis*. New York: Chapman and Hall ltd.

Figueras, M. (2000*). Introducción al análisis multivariante.* [Página Web en línea]. Recuperado el 22 de Diciembre de 2010, de http://www.5campus.com/leccion/anamul

Gondar, J. (2000). *Análisis factorial.* Artículos Estadísticos. Página Web en línea. Recuperado el 9 de Enero de 2010, de http: //www.estadistico.com

Malhotra, N. (1998*). Investigación de mercados: un enfoque práctico*. (Trad. V. de Parres). México D.F: Prentice Hall (Original en inglés).

Miquel, S.; Bigné E.; Levy, J. Y Miquel, A. (1997*). Investigación de mercados*. Madrid: Mc Graw-Hill/ Interamericana de España, S.A.

Richarm, M. (1997). *Eleven multivariate analysis techniques: key tool in your Marketing Research Survival Kit.*[Página Web en línea]. Recuperado el 23 de Diciembre de 2010, de http://www.decisionanalyst.com/pub-art/ Multivariate.asp

Schuschny, A y Soto, H. (2009). *Guía metodológica: diseño de indicadores compuestos de desarrollo sostenible*. CEPAL. División de Desarrollo Sostenible y Asentamientos Humanos. Santiago de Chile: Naciones Unidas.

Zikmund, W. (1998*). Investigación de mercados*. México D.F: Prentice Hall.